CANADIAN JOURNAL OF MATHEMATICS

Journal Canadien de Mathématiques

VOL. II - NO. 1

1950

Published for

THE CANADIAN MATHEMATICAL CONGRESS

by the

University of Toronto Press

The chief languages of the *Journal* are English and French.

Manuscripts for publication in the *Journal* should be sent to the *Editor-in-Chief*, H. S. M. Coxeter, University of Toronto. Every paper should contain an introduction summarizing the results as far as possible in such a way as to be understood by the non-expert.

All other correspondence should be addressed to the *Managing Editor*, G. de B. Robinson, University of Toronto.

The *Journal* is published quarterly. Subscriptions should be sent to the *Managing Editor*. The price per volume of four numbers is $6.00. This is reduced to $3.00 for individuals who are members of the following Societies:

Canadian Mathematical Congress
American Mathematical Society
Mathematical Association of America
London Mathematical Society
Société Mathématique de France

The Canadian Mathematical Congress gratefully acknowledges the assistance of the following towards the cost of publishing this Journal:

University of British Columbia
Loyola College
McGill University
Université de Montréal
Royal Military College
Ecole Polytechnique
University of Manitoba
McMaster University
Queen's University
University of Toronto
National Research Council
and the
American Mathematical Society

AUTHORIZED AS SECOND CLASS MAIL, POST OFFICE DEPARTMENT, OTTAWA

OLD AND NEW RESULTS ON KNOTS

HERBERT SEIFERT AND WILLIAM THRELFALL

THE theory of knots undertakes the task of giving a complete survey of all existing knots. A solid mathematical foundation was not laid to this theory until our century. A mathematician of the rank of Felix Klein thought it to be nearly hopeless to treat knot problems with the same exactness as we are accustomed to from classical mathematics. We want to give here a short summary of the modern topological methods enabling us to approach the knot problem in a mathematical way.

In order to exclude pathological knots, as for instance knots being entangled an infinite number of times, we will define a knot as a polygon lying in the space. In other words: a knot is a closed sequence of segments without double points. In Figure 1 some examples of knots are given in plane projection.

Now the question arises when two knots are to be called equivalent. One might be induced to call them so if one of them can be transformed into the other by a deformation without self-intersection. But this definition needs a restriction, otherwise every two knots would be equivalent. For one could transform both of them into an unknotted curve, a process shown for the trefoil knot by Figure 2. We therefore permit only more special transformations which do not allow such a tightening. We will call two knots *equivalent*, or of the same type, if they can be transformed into one another by a finite number of operations of the following kind: Let Δ be a triangle having one or two sides (and no other points) in common with the knot; then we add the boundary of Δ modulo 2 to the knot. This means the sides of Δ are to be added to the segments of the knot, and the segments, then occurring twice, are to be dropped. The two possible cases of such transformations are illustrated by Figure 3 and Figure 4. One may picture the knot as being pulled over the surface of the triangle.

The following definition has the same meaning, as can be proved: Two knots are equivalent if and only if the one can be transformed into the other by a topological, simplicial, sense-preserving mapping of the whole space onto itself [6]. These so-called semilinear self-transformations of space forming a group, the theory of knots may be regarded as part of the geometry belonging to this group.

We usually represent a knot in the drawing plane by parallel projection. The type of the knot is uniquely determined by its projection. It is always possible to choose the direction of the projecting lines in such a way that only "ordinary double points" occur, which means that in a double point one

Received September 1, 1948. Professor Threlfall died April 4, 1949.

Fig.1

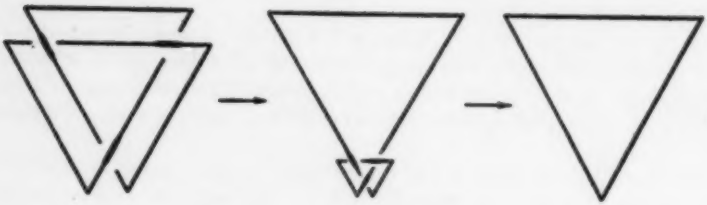


Fig.2

Fig.3

Fig.4

Fig.5

Fig.6

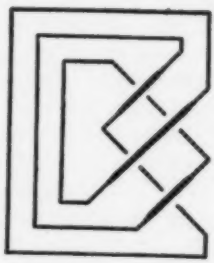


Fig.7

segment of the knot is crossed by one other segment only. The lower segment is drawn interrupted in the figures. One and the same knot and its equivalents are capable of an infinity of different projections. For instance Figures 5-7 are all projections of the trefoil knot.

There is one way near at hand for getting a survey of all possible knots. One has to construct systematically all projections with 2, 3, 4, . . . double points, and one has to find out by trying which of these projections represent equivalent knots. Of all the possible projections of the same type of knot, one distinguished by having the least number of double points or other simple qualities, will be picked out as a representative and be admitted to an inventory of knots. This is how the index of knots of Alexander and Briggs [2] was constructed. It contains 84 knots with up to 9 double points.

It remains to be seen whether different knots of the index are really not equivalent. This question, the true knot problem, cannot be decided by trying. For there is an infinity of possibilities of transforming a knot, and the reason why we may fail in transforming one knot of the index into another may be lack of skill or perseverance.

In order to prove two such knots to be really not equivalent, deeper methods are wanted. They offer themselves in the topological invariants of the complement of the knot, i.e., the space from which the knot has been taken away: if the knot k can be transformed into the knot k' by a semilinear transformation of the space R then the complements $R-k$ and $R-k'$ must be homeomorphic. Therefore the *topological invariants of $R-k$ are knot-invariants*. Thus the theory of knots is closely connected with the topology of three-dimensional manifolds, and every new topological invariant of three-dimensional manifolds is at the same time a new knot invariant. The only problem left to the knot theory is then to develop a method of calculating these invariants out of a given projection of the knot. Let us review the main results which have been attained in this direction.

One of the most important knot invariants is the *group of the knot* [4]. It is the fundamental group of the complement $R-k$. The fundamental group is defined for every connected complex L of any dimension. One has to choose a point O of L and to draw all closed oriented paths starting from O and running on L. Two such paths W and W' are called homotopic and are considered to be in the same class of paths, if W can be deformed on L into W', O remaining fixed, yet self-intersections being allowed throughout the deformation. The classes of paths are considered as the elements of a group; this is the fundamental group of L. The product W_1W_2 of two paths W_1 and W_2 is obtained by passing first along W_1 and then along W_2.

In the case of the group of a knot k, the complex L is the complement $R-k$. Thus the knot group of a "circle" is the free group of one generator; for every closed path that does not meet k can be deformed without intersecting k into a definite power of the path S entangling the circle once (Figure 8). The knot group therefore consists of all the powers of the class of the path S. The path

Fig. 8

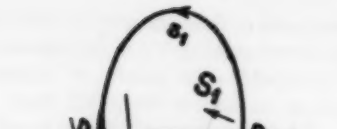


Fig. 9

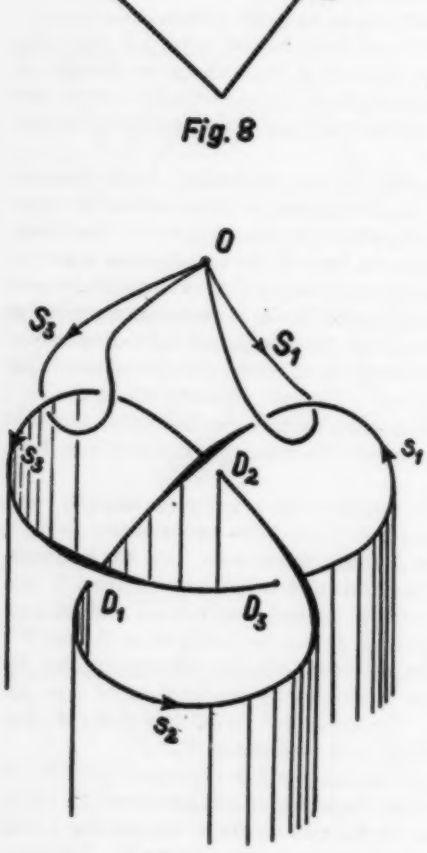


Fig. 10

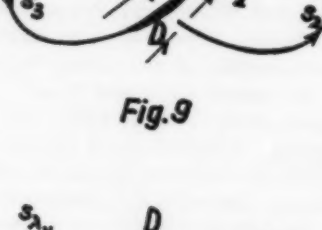


Fig. 11

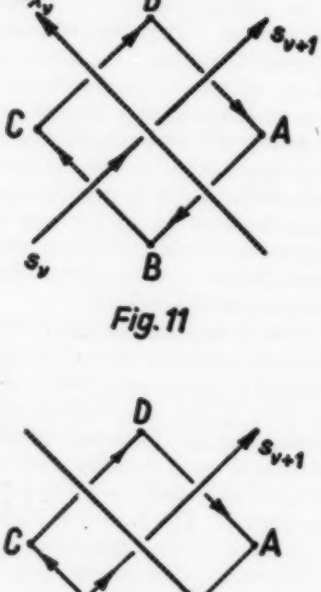


Fig. 12

W, for instance, shown in Figure 8, can be deformed into S^{-2}; the path E is homotopic to zero and represents the unity of the knot group.

Let k be an arbitrary knot given by its projection. We may represent its group by generators and defining relations in the following way. To every double point of the projection there correspond two points of the knot, one lying on the crossing, the other below, on the crossed branch. The latter may be called a crossed point. If there are n double points in the projection, then there will be n crossed points on the knot. We shall denote them by $D_1, D_2, \ldots, D_n$ in such an order as is given by a certain orientation of the knot. The knot is divided by the crossed points into n oriented segments $s_1, s_2, \ldots, s_n$, the segment s_ν running from $D_{\nu-1}$ to D_ν (indices are to be reduced modulo n so that D_0 means the same as D_n). We choose the origin O of the closed paths above the drawing plane, and we adjoin to the n segments $s_1, s_2, \ldots, s_n$ n elements $S_1, S_2, \ldots, S_n$ of the knot group which will generate the whole group. They are classes of paths of $R-k$. The class S_ν is represented by a path starting from O, entangling s_ν once in the positive sense and returning to O. To entangle in the positive sense means that the sense of the rotation of the oriented path around s_ν together with the orientation of s_ν determines a right-handed screw. We indicate such a way in the knot projection simply by an arrow, representing the part of the path which is overcrossed by the knot. Without changing the class of a path one may draw the path along the crossing segment over a double point. Figure 9 shows an example. Every class S_ν of paths is represented by two different arrows marked with the same letter. Figure 10 gives a picture of the same knot. In order to improve the view we added a cylinder underneath, with its generating lines in the direction of the projecting parallels.

To every double point there corresponds a certain defining relation of the knot group. To realize this let us consider Figures 11 and 12. They show the two possible kinds of crossing (overcrossing from right to left and from left to right). The closed path $ABCD$ represented in the projection is obviously homotopic to zero as it lies wholly beneath the knot. Now let the origin O of the closed paths be situated above the drawing-plane. If we then move the points A, B, C, D towards O, drawing the path $ABCD$ behind, this closed path will become a product of our generating elements S_μ. In the case of Figure 11 this product is

$$S_{\lambda_\nu} S_\nu S_{\lambda_\nu}^{-1} S_{\nu+1}^{-1},$$

and in the case of Figure 12 it is

$$S_{\lambda_\nu}^{-1} S_\nu S_{\lambda_\nu} S_{\nu+1}^{-1}.$$

We therefore have in the two cases the respective relations

(1a) $$R_\nu \equiv S_{\lambda_\nu} S_\nu S_{\lambda_\nu}^{-1} S_{\nu+1}^{-1} = 1,$$

(1b) $$R_\nu \equiv S_{\lambda_\nu}^{-1} S_\nu S_{\lambda_\nu} S_{\nu+1}^{-1} = 1.$$

These relations, formed for $\nu = 1, 2 \ldots, n$, constitute (as can be proved) a complete system of defining relations for the knot group. In the case of the trefoil knot (Figure 9) it is

$$(2) \qquad \begin{aligned} R_1 &\equiv S_2 S_1 S_3^{-1} S_2^{-1} = 1 \\ R_2 &\equiv S_1 S_2 S_1^{-1} S_3^{-1} = 1 \\ R_3 &\equiv S_2 S_3 S_2^{-1} S_1^{-1} = 1. \end{aligned}$$

It may be noticed, however, that the relations (1) are not all essential; one of them, being a consequence of the others, may be dropped.

We have so far constructed the knot group out of a given projection of the knot. But our result is still insufficient for distinguishing given knots. For the generating elements and the defining relations depend upon the choice of the projection, and no method is known for determining whether two groups given by generating elements and defining relations are identical or not. The problem of isomorphism of groups is as unsolved as the problem of equivalence of knots. For instance, the knot group of the trefoil knot for which we have found the defining relations (2) may as well be given by two generating elements A and B and the one defining relation $A^2 = B^3$. It is not at all obvious how the generating elements A and B may be expressed by S_1, S_2, S_3. Nevertheless the knot group is one of the most important knot invariants as it is the starting point for other and calculable invariants.

Besides of the fundamental group F of a complex L there is another well known invariant, the *homology group* H. One may define it as the abelianised fundamental group, that is the quotient group of F by its commutator group F_0:

$$H = F/F_0.$$

We see that the homology group can be derived from the fundamental group, and therefore it is in general a weaker invariant than the fundamental group. On the other hand it has the advantage of being determined by a finite system of numerical invariants which can be calculated by rational methods. For, every abelian group having a finite number of generating elements is the direct product of p cyclic groups of infinite order and certain cyclic groups of finite order, as is shown in the theory of elementary divisors. If the abelian group is the homology group H, the orders of the finite groups are called coefficients of torsion and the number p is the Betti number of the complex L.

Calculating the homology group H of the complement $L = R - k$ of a knot k one will find always the same result. H is the free group of one generator; in other words: $p = 1$ and there are no coefficients of torsion. This is a consequence of the relations (1). They reduce to $\overline{S}_\nu = \overline{S}_{\nu+1}$ by abelianising (the bar indicates the corresponding element of the abelianised group). The generators of the abelian group $\overline{S}_1, \overline{S}_2, \ldots, \overline{S}_n$ are therefore all equal to one another, say $= S$, and the abelianised knot group is the free cyclic group of one generator, S. It therefore cannot be utilized for the classification of knots.

But new invariants may be deduced from the *covering manifolds* of $R - k$. These covering manifolds are likewise connected with k in an invariant manner.

Alexander discovered that the homology groups of the covering manifolds are in general different for different knots. They may be therefore utilized for distinguishing knots. The so-called *cyclic* covering manifolds are of special importance. In order to construct them let us remark that an orientable surface without singularities may be framed in any knot k in such a way that k is the only boundary curve of the surface. In the case of the circle one may take an element (a 2-cell) (Figure 13); in the case of the trefoil knot the surface given by Figure 14 will do. This figure shows the surface bounded by the knot in plane projection. In the three-dimensional space the two hatched parts of the projection cohere along three segments, which are double points in the projection. This surface is a perforated torus as may be shown by calculating its Euler characteristic. The surface of Figure 15, however, would not be fit for our purpose, as it is non-orientable (a Möbius strip).

We now cut the space R along the surface. We get a three-dimensional "sheet," and we attach g replicas of this sheet to one another in cyclic order around the knot. The analogous process, one dimension lower, is the construction of a Riemann surface on the sphere: the sphere is cut along an arc; it becomes a two-dimensional sheet, and g of these sheets are to be attached around the two branch points. In this way we get a g-fold covering manifold R_g of R, the knot k being the branch-line. By taking k out of R_g we obtain a covering manifold $R_g - k$ "*without* ramification." It is to be noticed that R_g depends only on the number g and the knot k, but not on the surface used for the construction. (However, there exist in general still other non-cyclic covering manifolds which are not to be considered here.)

In order to calculate the fundamental group F_g and the homology group H_g of $R_g - k$ we proceed from the theorem that the fundamental group of a covering manifold is isomorphic with a subgroup of the fundamental group F of the basic manifold. We obtain the subgroup by copying through into the basic manifold all oriented closed paths of the covering manifold starting from a fixed point. We therefore have to find in $R - k$ those closed paths W starting from O to which correspond in $R_g - k$ closed paths starting from one and the same point O. These are exactly the paths whose intersection number with the surface bounded by k is a multiple of g. An intersection point is to be counted positive or negative according as W pierces the surface from right to left or from left to right. (We may speak of left and right as our surface is orientable and therefore two-sided in the space). The intersection number is called also the *looping coefficient* of W with k. It is independent of the surface. We thus have found: the fundamental group F_g of $R_g - k$ is isomorphic with the subgroup of the (classes of) paths, the looping coefficients of which with k are multiples of g.

Reidemeister [7] has shown how to derive generators and defining relations of a subgroup from generators and defining relations of the whole group. By applying this method to the knot group one arrives after some calculations [13] at the following simple result.

Fig.13

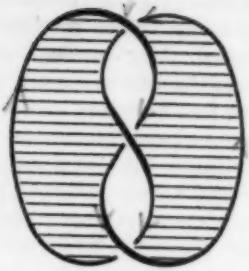


Fig.14

Fig.15

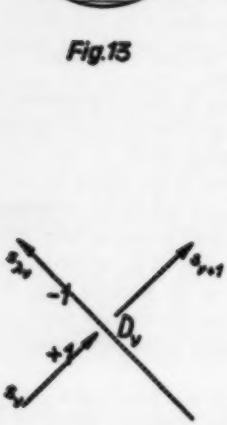


Fig.16

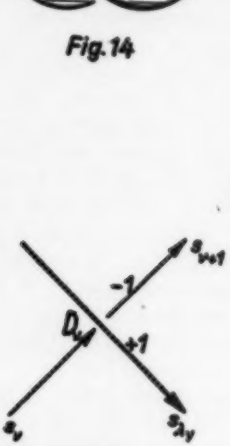


Fig.17

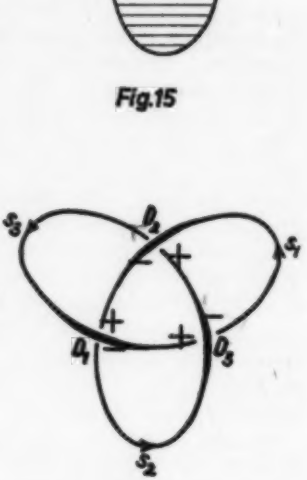


Fig.18

Fig.19

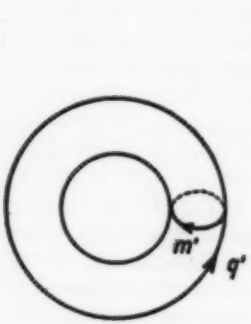


Fig.20

In order to determine the homology group H_g of the g-fold cyclic covering manifold $R_g - k$, the double points of the knot projection are to be denoted seriatim by $D_1, D_2, \ldots, D_n$, and the corresponding segment running from one crossed point to the next by $s_1, s_2, \ldots, s_n$ as has been done above. Let s_{λ_ν} be the crossing segment in the double point D_ν. Then we will write on this segment at the point D_ν the number -1 or $+1$ according as the crossing takes place from right to left or from left to right. We attach to the segment lying on the left side of s_{λ_ν} and coinciding with D_ν the opposite number, $+1$ or -1 as shown in Figures 16 and 17. From these numbers we form the matrix A, the rows corresponding to the double points $D_1, D_2, \ldots, D_n$ and the columns to the segments $s_1, s_2, \ldots, s_n$. (It may happen that in the case of a crossing from right to left s_ν is the same as s_{λ_ν}. Then s_ν will have two numbers $+1$ and -1 as suffixes of D_ν. In this case we put the sum of these two numbers, i.e., 0, in the position (D_ν, s_ν) of the matrix A; similarly in the case of a crossing from left to right.) We obtain from this matrix by suppressing the last row and the last column a matrix $\bar{A}$, and from this by adding the first column to the second, then the second to the third, etc., a matrix Γ of $n-1$ rows and $n-1$ columns. The torsion coefficients of $R_g - k$ are then the elementary divisors (other than 1) of the matrix

$$\Gamma^g - (\Gamma - E)^g,$$

E being the unit matrix of $n-1$ rows, whereas the rank defect of this matrix, augmented by one, gives the Betti number of $R_g - k$.

This theorem allows us to calculate the homology groups of all cyclic covering manifolds by means of one and the same matrix Γ. As Alexander has shown, one may thus verify that the 84 knots of the Alexander-Briggs index are distinct with a few exceptions.

Let us take as an example the trefoil knot. According to our prescription we number the double points and the segments, and we determine the indices ± 1. In Figure 18 we have written the signs $+$ and $-$ of the indices on the segments. The matrices A, $\bar{A}$ and Γ become

$$A = \begin{array}{c|ccc} & s_1 & s_2 & s_3 \\ \hline D_1 & 1 & 0 & -1 \\ D_2 & -1 & 1 & 0 \\ D_3 & 0 & -1 & 1 \end{array}, \quad \bar{A} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}.$$

We therefore get

$$\Gamma^2 - (\Gamma - E)^2 = \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix}.$$

By suitable transformations of the rows and columns the normal form $\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$ results. It has one elementary divisor 3, and the rank defect is 0. Thus the

twofold cyclic covering manifold $R_2 - k$ has one torsion coefficient of value 3 and the Betti number 1. In the same way one may calculate

$$\Gamma^3 - (\Gamma - E)^3 = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix},$$

that is to say, the threefold covering manifold $R_3 - k$ has two torsion coefficients of value 2 each, and the Betti number 1. For $R_4 - k$ we find

$$\Gamma^4 - (\Gamma - E)^4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

This covering manifold has therefore no torsion coefficient, and the Betti number is 3.

Next to the covering manifolds $R_q - k$ of a finite number of sheets, the infinite cyclic covering manifold $R_0 - k$ is important. Its fundamental group is isomorphic with the subgroup F_0 of those elements of the knot group, whose looping coefficients with k are 0. (F_0 is the commutator group of F.) Because of the infinite number of sheets, the homology group H_0 of $R_0 - k$ can in general be given only by an infinite number of generators and relations. But a finite representation of H_0 may be obtained by interpreting H_0 as a group with operators in the following sense. There is a symmetry operation (Deckbewegung) of $R_0 - k$ transferring every sheet into the next one. It induces in H_0 an automorphism x. By making use of the operator x one obtains instead of the infinite number of generators and relations of H_0 a finite number of relations, the domain of coefficients being the integral domain of the polynomials in x and x^{-1} with integral coefficients.

It can be shown that the polynomial

$$\Delta = \left| \Gamma - x(E - \Gamma) \right|,$$

Γ being the above-mentioned matrix, is a knot invariant, if one does not count a factor $\pm x^p$ which remains undetermined. This is the "*L-polynomial*" introduced by Alexander [1], see also Seifert [9]. In the case of the trefoil knot it is

$$\Delta = 1 + x + x^2,$$

whereas the circle has the L-polynomial

$$\Delta = 1.$$

An interesting application of the L-polynomials is the following relation between the L-polynomials of a special class of knots. Let k be an (oriented) knot lying in the space R, V a tubular neighbourhood of k. The boundary of V is a (two-dimensional) torus T. Let $W = R - V + T$ be the closed complement of V in R. A closed oriented curve on T without double points and non-bounding on T is called a "meridian" of V if it bounds on V, and a "parallel" of V if it bounds on W.

V can be mapped by a topological representation ϕ on an unknotted solid tube V', bounded by a two-dimensional torus T', in such a way that the parallel q of V becomes a parallel q' of V'. Figures 19 and 20 illustrate the case of k being the trefoil knot.

Now let l be an arbitrary knot lying in V. l being a closed 1-chain of V, it is homologous on V to a multiple of k, say

$$l \sim nk \text{ on } V.$$

We may assume $n \geqq 0$ by orienting l properly. The topological representation ϕ of V on V' maps l into a knot l' of V'. Let $\Delta_k(x)$, $\Delta_l(x)$, $\Delta_{l'}(x)$ be the L-polynomials of the knots k, l, l' respectively. Then our theorem is (Seifert [12])

$$\Delta_l(x) = \Delta_{l'}(x)\Delta_k(x^n).$$

This theorem contains a result of Burau [3] concerning the special case where l is a "tube knot," the carrier knot of which is k.

An example of our theorem is given in Figure 21. Here the knot k is the trefoil knot, and $n = 0$. The case $n = 0$ has a remarkable consequence. Then we have $\Delta_k(x^n) = \Delta_k(x^0) = \Delta_k(1)$. But $\Delta_k(1) = 1$ for every knot. Therefore for $n = 0$ our theorem is $\Delta_l(x) = \Delta_{l'}(x)$. In other words the L-polynomial of l does not depend on the knot k. The doubled knots in the sense of Whitehead [14] are a special case hereof. We shall treat them soon again.

Besides the homology groups there are still other invariants of three-dimensional manifolds which play a role in distinguishing knots. In three-dimensional space, two closed oriented curves without common points have a certain looping coefficient. In the same way a looping coefficient of two closed oriented curves may be defined in other orientable three-dimensional manifolds, provided that the curves themselves or multiples of them be homologous to zero. These looping coefficients, however, will be in general *fractions*. For instance the looping coefficient of two projective lines of the projective space is 1/2. The looping coefficients possess the important property of changing their sign when the orientation of the space is reversed. From them may be deduced the so-called *looping invariants*, which are invariants of a three-dimensional manifold including a certain orientation. They may change when the orientation changes. There are examples where they do more for distinguishing knots than the knot group. For instance, it can be shown with them that the two knots of Figures 22 and 23 are not equivalent in spite of their having the same knot group. They further allow us in many cases to distinguish a knot from its symmetric (its image in a mirror). This is, for instance, the case with the trefoil knot, and the result is not to be had by the homology groups, H_0, H_2, H_3, ... It may be mentioned that these invariants are connected with the so-called quadratic form of the knot ([5] and [10]).

We have hitherto enumerated the most important knot invariants so far as they can be defined for *arbitrary* knots. What is the significance of these invariants?

The problem of isomorphism of the knot group is unsolved. Setting aside the knot group, we have the homology groups and the looping invariants of the infinite number of covering manifolds $R_\varrho - k$. Let us call them the *homology invariants* of the knot. They suffice for distinguishing all the 84 knots of

Fig.21

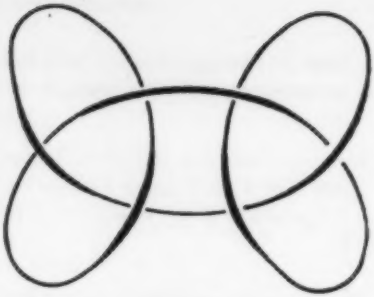


Fig.22

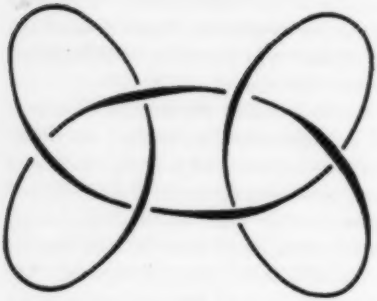


Fig.23

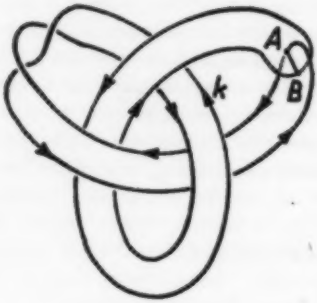


Fig.24

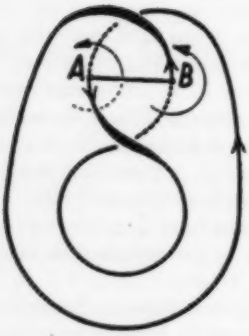


Fig.25

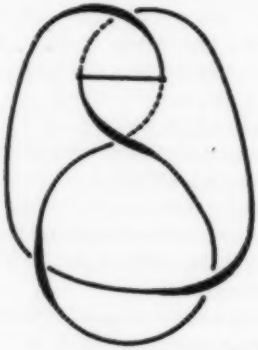


Fig.26

the Alexander-Briggs index. But the knot problem is still far from being solved with them. For there exists an infinite number of knots all having the same homology invariants.

There is even an infinite number of knots having the same homology invariants as the circle. For instance, certain *doubled knots* of J. H. C. Whitehead [14] are of this kind. They are defined as follows. Take a narrow ribbon, knot it in an arbitrary way and after that put the ends one into the other, so that they will penetrate each other along a segment AB, as is shown in Figure 24. Then the boundary of the ribbon will be a doubled knot. By making the ribbon more and more narrow it will finally be reduced to a knotted line $\bar{k}$, which we call the carrier knot. The carrier knot of the doubled knot shown in Figure 24 is the trefoil knot. To every carrier knot $\bar{k}$ there corresponds an infinite number of doubled knots. To construct them we only have to knot the ribbon according to $\bar{k}$ and then to twist it an arbitrary number of times. By choosing this number suitably the homology invariants of the doubled knot will be the same as those of the circle.

How is it possible to distinguish such knots? Whitehead has proved that two doubled knots can be equivalent only if their carrier knots have the same knot group. We can prove in addition that not only the knot group of $\bar{k}$ but the type of $\bar{k}$ itself is an invariant of the doubled knot. Two doubled knots, if constructed by means of inequivalent carrier knots, are therefore certainly not equivalent.

In contrast to the above-mentioned homology invariants, which may be called *algebraic* topological invariants, the invariance of the carrier knot is of a purely topological nature. Accordingly, other methods have to be used to prove it. One may proceed as follows. By definition a doubled knot is the boundary of an "element with self-penetration." This means a 2-cell of the three-dimensional space, the only singularity of which is a self-penetration along a segment AB (Figure 24). Yet we have to pay attention to the fact that this element with self-penetration is not determined by the doubled knot. For instance, one may deform it into another one without changing the boundary.

The question therefore arises how many essentially different elements with self-penetration may be bounded by the same doubled knot k. Let us explain this question as applied to the simplest example imaginable. The circle itself may be interpreted as a doubled knot, as is seen by Figure 25. The figure shows the ribbon bounded by the circle. We may provide the ribbon with a certain "index." It is defined as follows. We give the boundary k of the ribbon a certain orientation. This induces a certain orientation of the ribbon itself. A and B being the points of penetration between k and the ribbon, the orientation of k together with the orientation of the ribbon determines a certain screw (space orientation) in A. In Figure 25 this is a right-handed screw. The same right-handed screw is determined in B.

If we reverse the orientation of k, the orientation of the ribbon is reversed at the same time. Therefore the sense of the screws does not change. Now we assign to the intersection point A the index $+1$ if the screw of A is a right-handed screw, and -1 if it is a left-handed screw, and we do the same to the other point B. Then the index of the ribbon is the sum of the two indices of A and B. In the case of Figure 25 the index is $+2$. We therefore see that a circle may bound a ribbon of index $+2$. But then it may bound a ribbon of index -2 just as well. We only have to reflect Figure 25. This process reverses the orientation of space and therefore right-handed screws become left-handed screws. The reflected knot is again a circle, but now it bounds an element with self-penetration of index -2. The result is, that a circle can bound two essentially different elements with self-penetration, which cannot be transformed into one another by a semilinear mapping of the space, for such a mapping would not change the index of the element.

We find the same situation in the case of the "four-knot." The ribbon shown in Figure 26 has the index $+2$. By reflecting it we get a ribbon of index -2. Now it is known that the reflected image of the four-knot is equivalent to the original knot. Therefore the four-knot also admits two essentially different elements with self-penetration of which it is the boundary.

We believe the circle and the four-knot to be the only two knots having this quality. We can prove the following result: *If a doubled knot k can be constructed by means of a carrier knot $\bar{k}$, $\bar{k}$ not being the circle, then there exists, apart from semilinear mappings of the space, only one element with self-penetration bounded by k.* The proof of this theorem is rather complicated. The only way is to construct the semilinear mapping of one of two elements with self-penetration bounded by k into the other. This can be done by considering the lines and points of mutual penetration of the two elements and splitting them off one by one by appropriate methods [11].

The invariance of the carrier knot is an immediate consequence of this theorem. Given the element with self-penetration, the carrier knot is obtained by joining the ends A and B of the penetrating segment by a curve running along the ribbon. The resulting closed curve is itself the carrier knot $\bar{k}$. Now let k and k' be two doubled knots derived from the carrier knots $\bar{k}$ and $\bar{k}'$. Then the equivalence of $\bar{k}$ and $\bar{k}'$ follows from the equivalence of k and k' provided that one at least of the two knots $\bar{k}$ and $\bar{k}'$ is not the circle. In the case when both of them are circles, they are obviously equivalent. In any case, therefore, the carrier knot is an invariant of the doubled knot.

Furthermore it follows from our theorem of the uniqueness of the bounded element that a doubled knot, the carrier knot of which is not a circle, will never be equivalent to its mirrored image. For the index of the bounded element would change by reflection. If the doubled knot were equivalent to its image it would bound therefore two essentially different elements of index $+2$ and -2.

The preceding theorems are concerned with special classes of knots. Let us conclude with a result relating to the decomposition of an arbitrary knot. We shall call a knot a prime knot if it is impossible to cut it after a suitable deformation by a plane having only two points P and Q in common with k, into two knots (both distinct from the circle) consisting of the two parts of k and the closing segment PQ. For instance the knot shown in Figure 22 can be decomposed into two trefoil knots. Now our theorem is, that every knot k can be decomposed into prime knots and that the series of these prime knots is unique but for the order of them [8].

References

[1] J. W. Alexander, "Topological Invariants of Knots and Links," *Trans. Amer. Math. Soc.*, vol. 30 (1928), 275-306.

[2] J. W. Alexander and C. Briggs, "On Types of Knotted Curves," *Ann. of Math.*, vol. 28 (1926/27), 562-586.

[3] W. Burau, "Kennzeichnung der Schlauchknoten," *Abh. Math. Sem. Hamburg Univ.*, vol. 9 (1932), 125-133.

[4] M. Dehn, "Über die Topologie des dreidimensionalen Raumes," *Math. Ann.*, vol. 69 (1910), 137-168; "Die beiden Kleeblattschlingen," *Math. Ann.*, vol. 75 (1914).

[5] L. Goeritz, "Knoten und quadratische Formen," *Math. Zeit.*, vol. 36 (1933), 647-654.

[6] W. Graeub, *Semilineare Abbildungen*—to appear in *Sitz. Ber. Ak. d. Wissensch.*, Heidelberg.

[7] K. Reidemeister, *Knotentheorie*, Berlin, 1932 (*Erg. d. Math.*, vol. 1).

[8] H. Schubert, "Die eindeutige Zerlegbarkeit eines Knotens in Primknoten," *Sitz. Ber. Ak. d. Wissensch.*, Heidelberg (1949).

[9] H. Seifert, "Über das Geschlecht von Knoten," *Math. Ann.*, vol. 110 (1934).

[10] H. Seifert, "Die Verschlingungsinvarianten der cyklischen Knotenüberlagerungen," *Abh. Math. Sem. Hamburg Univ.*, vol. 11 (1935).

[11] H. Seifert, "Schlingknoten," *Math. Zeit.*, vol. 52 (1949), 62-80.

[12] H. Seifert, *Über die L-Polynome einer speziellen Klasse von Knoten*—to appear in *Quart. J. Math.*

[13] W. Threlfall, *Knotengruppe und Homologieinvarianten*—to appear in *Sitz. Ber. Ak. d. Wissensch.*, Heidelberg.

[14] J. H. C. Whitehead, "On Doubled Knots," *London Math. Soc.*, vol. 12 (1937), 63-71.

ON THE FIELD OF ORIGIN OF AN IDEAL

H. B. MANN

In this paper we shall consider integral ideals in finite algebraic extensions $(\mathfrak{F}, \mathfrak{F}_1, \ldots)$ of the field of rational numbers.

Two ideals $\mathfrak{a}$, $\mathfrak{b}$ in the same field $\mathfrak{F}$ are said to be equal if and only if they contain the same numbers.

Let $\mathfrak{F}_1 \supset \mathfrak{F}_2$ and let $\mathfrak{A}$ be an ideal in $\mathfrak{F}_2$. The numbers of $\mathfrak{A}$ generate an ideal $\mathfrak{a}$ in $\mathfrak{F}_1$ and it is known that the intersection $\mathfrak{a} \cap \mathfrak{F}_2 = \mathfrak{A}$. (See for instance Hecke, *Theorie der algebraischen Zahlen*, § 37). Also if $\mathfrak{a} \subset \mathfrak{F}_1$ and $\mathfrak{b} \subset \mathfrak{F}_2$ generate the same ideal in a field containing $\mathfrak{F}_1$ and $\mathfrak{F}_2$ then they must generate the same ideal in $\mathfrak{F}_1 \cup \mathfrak{F}_2$ and thus in every field containing $\mathfrak{F}_1$ and $\mathfrak{F}_2$.

We shall therefore call two ideals $\mathfrak{a}$ and $\mathfrak{b}$ equal if they generate the same ideal in a field containing all the numbers of $\mathfrak{a}$ and of $\mathfrak{b}$. Two such ideals may therefore be denoted by the same symbol and we shall speak of an ideal $\mathfrak{a}$ without regard to a particular field. An ideal $\mathfrak{a}$ will be said to be contained in a field $\mathfrak{F}$ if it may be generated by numbers in $\mathfrak{F}$; in other words, if it has a basis in $\mathfrak{F}$.

It seems natural to try to characterize those fields which contain a given ideal $\mathfrak{a}$, and in this paper we shall find such a characterization at least in the case that a power of $\mathfrak{a}$ is a prime ideal in some extension of $\mathfrak{F}$.

A necessary and sufficient condition for an ideal $\mathfrak{a}$ to be contained in a given field $\mathfrak{F}$ will be derived in the case that $\mathfrak{a}$ is an ideal of order 1, as defined in this paper. For prime ideals of order greater than 1 a necessary and sufficient condition will also be given.

From now on we shall consider finite algebraic extensions $(\mathfrak{F}_1, \ldots)$ over a field $\mathfrak{F}_1$ itself a finite algebraic extension over the field of rational numbers. Admissible subfields of $\mathfrak{F}_1$ are those containing $\mathfrak{F}$. Throughout the paper only fields containing $\mathfrak{F}$ will be considered.

Consider an ideal $\mathfrak{a} \subset \mathfrak{F}_1$. Either $\mathfrak{a}$ is not contained in any admissible subfield of $\mathfrak{F}_1$ or $\mathfrak{F}_1$ must contain an admissible subfield $\mathfrak{F}_2$ which has the property that $\mathfrak{a}$ is in $\mathfrak{F}_2$ but not in any admissible subfield of $\mathfrak{F}_2$. We therefore define:

Definition 1. *If $\mathfrak{a}$ is in $\mathfrak{F}_1$ but not in any proper admissible subfield of $\mathfrak{F}_1$ then $\mathfrak{a}$ is said to originate in $\mathfrak{F}_1$ over $\mathfrak{F}$.*

Consider $\mathfrak{F}_1 \supset \mathfrak{F}_2$ and let $\mathfrak{a}$ be an ideal in $\mathfrak{F}_1$. The numbers of $\mathfrak{a}$ which lie in $\mathfrak{F}_2$ form an ideal $\mathfrak{A}$ in $\mathfrak{F}_2$. This ideal $\mathfrak{A}$ is said to correspond in $\mathfrak{F}_2$ to the ideal $\mathfrak{a}$. The ideal $\mathfrak{A}$ depends only on $\mathfrak{a}$ but not on $\mathfrak{F}_1$.

Definition 2. If $\mathfrak{A} \subset \mathfrak{F}$ *corresponds to $\mathfrak{a}$ in $\mathfrak{F}_1$ and*

$$\mathfrak{A} = \mathfrak{a}^e \mathfrak{c}, \quad (\mathfrak{a}, \mathfrak{c}) = 1 \tag{1}$$

then $\mathfrak{a}$ is said to be of order e with respect to $\mathfrak{F}$.

Received September 1, 1948.

REMARK. Not every ideal has an order with respect to $\mathfrak{F}$; however, every ideal which is a prime ideal in some extension of $\mathfrak{F}$ does.

THEOREM 1. *If $\mathfrak{a}$ is an ideal of order 1 with respect to $\mathfrak{F}$ then $\mathfrak{a}$ originates in a unique subfield $\mathfrak{F}_1$ over $\mathfrak{F}$. An extension $\mathfrak{F}' \supset \mathfrak{F}$ contains $\mathfrak{a}$ if and only if it contains $\mathfrak{F}_1$.*

Proof. If $\mathfrak{a}$ does not originate in $\mathfrak{F}'$, then it must originate in some subfield of $\mathfrak{F}'$. Hence $\mathfrak{a}$ originates in at least one field.

Suppose then that $\mathfrak{a}$ originates in $\mathfrak{F}_1$ and also in $\mathfrak{F}_2$. Let $\mathfrak{F}_n$ be a normal extension of $\mathfrak{F}$ containing $\mathfrak{F}_1$ and $\mathfrak{F}_2$ and $\mathfrak{G}$ the Galois group of $\mathfrak{F}_n$ over $\mathfrak{F}$. Let $\mathfrak{H}_1$ and $\mathfrak{H}_2$ be the subgroups of $\mathfrak{G}$ leaving $\mathfrak{F}_1$ and $\mathfrak{F}_2$ respectively fixed. Since $\mathfrak{a}$ has a basis in $\mathfrak{F}_1$ and in $\mathfrak{F}_2$ it follows that $\mathfrak{a}$ is transformed into itself by the union $\mathfrak{H}_1 \cup \mathfrak{H}_2 = \overline{\mathfrak{H}}$. To $\overline{\mathfrak{H}}$ corresponds the field $\overline{\mathfrak{F}} = \mathfrak{F}_1 \cap \mathfrak{F}_2$ which certainly contains $\mathfrak{F}$. Let $\bar{\mathfrak{a}} \subset \overline{\mathfrak{F}}$ and $\mathfrak{A} \subset \mathfrak{F}$ correspond to $\mathfrak{a} \subset \mathfrak{F}_1$ then

$$\bar{\mathfrak{a}} = \mathfrak{a}\mathfrak{c}' \tag{2}$$
$$\mathfrak{A} = \bar{\mathfrak{a}}\mathfrak{b} = \mathfrak{a}\mathfrak{c}'\mathfrak{b}.$$

Since $\mathfrak{c}'\mathfrak{b} = \mathfrak{c}$ by (1) and since $(\mathfrak{c}, \mathfrak{a}) = 1$ by hypothesis we must have

$$(\mathfrak{c}', \mathfrak{a}) = 1. \tag{3}$$

If

$$\overline{\mathfrak{H}} = \mathfrak{H}_1 + \mathfrak{H}_1 A_2 + \ldots + \mathfrak{H}_1 A_g \tag{4}$$

then all relative conjugate fields of $\mathfrak{F}_1$ over $\overline{\mathfrak{F}}$ are obtained each once by applying $1, A_2, \ldots, A_g$ to $\mathfrak{F}_1$. Hence since A_i transforms $\mathfrak{a}$ into itself

$$\mathfrak{a} = \mathfrak{a}^{A_2} = \ldots = \mathfrak{a}^{A_g}. \tag{5}$$

Thus

$$\bar{\mathfrak{a}} = \mathfrak{a}\mathfrak{c}'^{A_i} \qquad (i = 1, \ldots, g), \tag{6}$$
$$\mathfrak{c}'^{A_i} = \mathfrak{c}'.$$

Thus

$$\mathfrak{a}^g \subset \overline{\mathfrak{F}}, \ \mathfrak{c}'^g \subset \overline{\mathfrak{F}}. \tag{7}$$

Since $\mathfrak{a}^g \subset \overline{\mathfrak{F}}$, we must have $\mathfrak{a}^g \subset \bar{\mathfrak{a}}$ and

$$\mathfrak{a}^g = \bar{\mathfrak{a}}\mathfrak{b}' = \mathfrak{a}\mathfrak{c}'\mathfrak{b}'. \tag{8}$$

Hence $\mathfrak{c}' = (1)$ since otherwise $(\mathfrak{a}, \mathfrak{c}') \neq 1$ contradicting (3). Thus by (2) $\mathfrak{a} = \bar{\mathfrak{a}}$ and since by hypothesis $\mathfrak{a}$ originates in $\mathfrak{F}_1$ and $\mathfrak{F}_2$ it follows that $\overline{\mathfrak{F}} = \mathfrak{F}_1 = \mathfrak{F}_2$.

If now $\mathfrak{a}$ is in $\mathfrak{F}'$ then $\mathfrak{F}'$ must contain a field in which $\mathfrak{a}$ originates. Hence $\mathfrak{F}'$ must contain $\mathfrak{F}_1$. Conversely if $\mathfrak{F}' \supset \mathfrak{F}_1$ then $\mathfrak{F}' \supset \mathfrak{a}$ since $\mathfrak{a} \subset \mathfrak{F}_1$.

THEOREM 2. *If $\mathfrak{p}$ is an ideal in any field over $\mathfrak{F}$ and g is the largest integer for which $\mathfrak{p}^g$ is a prime ideal in some extension of $\mathfrak{F}$ then $\mathfrak{p}^g$ originates in a unique extension $\mathfrak{F}' \supset \mathfrak{F}$ and is a prime ideal in $\mathfrak{F}'$. Moreover every field that contains a power of $\mathfrak{p}$ contains $\mathfrak{F}'$.*

Proof. Let $\mathfrak{P}$ in $\mathfrak{F}$ correspond to $\mathfrak{p}$. Since $\mathfrak{p}^g$ is a prime ideal in some field over $\mathfrak{F}$, $\mathfrak{P}$ must be a prime ideal. That is to say

$$\mathfrak{P} = \mathfrak{p}^e\mathfrak{a}, \quad (\mathfrak{p}, \mathfrak{a}) = 1. \tag{9}$$

Thus $\mathfrak{p}^e$ satisfies the conditions of Theorem 1. Let $\mathfrak{F}'$ be the unique field in which $\mathfrak{p}^e$ originates. Let $\mathfrak{p}^g$ be a prime ideal in $\mathfrak{F}''$. To $\mathfrak{p}^g$ corresponds a prime ideal in $\mathfrak{F}$ and since this prime ideal has a common factor with $\mathfrak{P}$ it must be equal to $\mathfrak{P}$. Thus since $(\mathfrak{p}, \mathfrak{a}) = 1$

$$\mathfrak{P} = (\mathfrak{p}^g)^t\mathfrak{a}, \quad e \equiv 0(g), (\mathfrak{p}^g, \mathfrak{a}) = 1. \tag{10}$$

Thus $\mathfrak{F}''$ contains $\mathfrak{p}^g$ hence must also contain $\mathfrak{F}'$. Moreover $\mathfrak{p}^g$ is a prime ideal in $\mathfrak{F}'$ since it is prime in $\mathfrak{F}''$ and since g is the largest power of $\mathfrak{p}$ which is prime in any field. Every field that contains a power of $\mathfrak{p}$ must contain $\mathfrak{p}^e$ hence must contain $\mathfrak{F}'$. In particular $\mathfrak{p}^g$ cannot be contained in any subfield of $\mathfrak{F}'$ and therefore originates in $\mathfrak{F}'$.

COROLLARY. *If $\mathfrak{p}$ is an ideal in some extension $\mathfrak{F}'$ of $\mathfrak{F}$ and $\mathfrak{p}^g$ is the highest power of $\mathfrak{p}$ which is a prime ideal in an admissible subfield of $\mathfrak{F}'$ then $\mathfrak{p}^g$ is the highest power of $\mathfrak{p}$ which is a prime ideal in any extension of $\mathfrak{F}$. (We may take $g = 0$ if no power of $\mathfrak{p}$ is a prime ideal in any admissible subfield of $\mathfrak{F}'$.)*

A simple example is the ideal $(\sqrt{2})$, when f is the field of rational numbers. Here $g = e = 2, f = f'$.

THEOREM 3. *If $\mathfrak{p}$ is a prime ideal in some extension of $\mathfrak{F}$ and $\mathfrak{p}^g$ is the largest power of $\mathfrak{p}$ which is a prime ideal of any extension of $\mathfrak{F}$ and if $\mathfrak{p}^h$ is a prime ideal in some extension $\mathfrak{F}_1$ of $\mathfrak{F}$ then*

$$g \equiv 0(h). \tag{11}$$

Let $\mathfrak{F}'$ be the unique field in which $\mathfrak{p}^g$ originates by Theorem 2. By the same theorem we have

$$\mathfrak{F}' \subset \mathfrak{F}_1. \tag{12}$$

To $\mathfrak{p}^h$ corresponds a prime ideal in $\mathfrak{F}'$ which has a common factor with $\mathfrak{p}^g$ and therefore must equal $\mathfrak{p}^g$ since $\mathfrak{p}^g$ is a prime ideal in $\mathfrak{F}'$. Thus

$$\mathfrak{p}^g = (\mathfrak{p}^h)^t, \quad g = ht. \tag{13}$$

If $\mathfrak{p}$ is a prime ideal in some extension of $\mathfrak{F}$ but no power of $\mathfrak{p}$ is a prime ideal in any extension of $\mathfrak{F}$ then by Theorem 2 there is a unique extension of $\mathfrak{F}$ in which $\mathfrak{p}$ originates over $\mathfrak{F}$. Quite in contrast to this we shall show that if $\mathfrak{p}^g$ $(g > 1)$ is a prime ideal in some extension of $\mathfrak{F}$ then there are infinitely many extensions of $\mathfrak{F}$ in which $\mathfrak{p}$ originates and is a prime ideal. We show this by proving

THEOREM 4. *If $\mathfrak{p}$ is a prime ideal in $\mathfrak{F}$ then for every $g > 1$ there exists an ideal $\mathfrak{P}$ such that $\mathfrak{P}^g = \mathfrak{p}$. The ideal $\mathfrak{P}$ originates as a prime ideal in infinitely many fields over $\mathfrak{F}$.*

Proof. Let $\mathfrak{p} = (\alpha_1, \alpha_2)$, $\alpha_1, \alpha_2 \subset \mathfrak{F}$. We may choose

$$(\alpha_2) = \mathfrak{p}\mathfrak{c}, \quad (\mathfrak{p}, \mathfrak{c}) = 1. \tag{14}$$

Choose q prime to a_1, a_2, $\mathfrak{p}$ and to the absolute differente of $\mathfrak{F}(\zeta)$, where ζ is a primitive gth root of unity, and square free. In $\mathfrak{F}(\sqrt[g]{qa_2})$ the ideal $\mathfrak{p}$ is the gth power of the ideal $\mathfrak{P} = (a_1, \sqrt[g]{qa_2})$, for a_1 and $\sqrt[g]{qa_2}$ can have only a divisor $\overline{\mathfrak{P}}$ of $\mathfrak{p}$ in common. Thus

$$\begin{aligned} a_1 &= \mathfrak{p}\mathfrak{A} \\ \sqrt[g]{qa_2} &= \overline{\mathfrak{P}}\mathfrak{B} \qquad (\mathfrak{p}, \mathfrak{B}) = 1 \\ qa_2 &= \overline{\mathfrak{P}}^g\mathfrak{B}^g = \mathfrak{p}\mathfrak{c}q, \quad (\mathfrak{p}, \mathfrak{c}) = 1, \quad \overline{\mathfrak{P}}^g = \mathfrak{p}. \end{aligned}$$

Hence $\mathfrak{P}^g = (a_1, \sqrt[g]{qa_2})^g = \overline{\mathfrak{P}}^g = \mathfrak{p}$.

We shall show now that $\mathfrak{F}(\sqrt[g]{qa_2}) \neq \mathfrak{F}(\sqrt[g]{q'a_2})$ if $(q) \neq (q')$. The numbers qa_2 and $q'a_2$ are square free in $\mathfrak{F}(\zeta)$ by assumption. Hence the polynomials $x^g - qa_2$, $x^g - q'a_2$ are irreducible in $\mathfrak{F}(\zeta)$ by Eisenstein's criterion. Thus 1, $\sqrt[g]{qa_2}, \ldots, (\sqrt[g]{qa_2})^{g-1}$ are independent over $\mathfrak{F}(\zeta)$. If $\sqrt[g]{q'a_2} \subset \mathfrak{F}(\sqrt[g]{qa_2})$ then

$$\sqrt[g]{q'a_2} = a_0 + a_1\sqrt[g]{qa_2} + \ldots + a_{g-1}(\sqrt[g]{qa_2})^{g-1}$$

applying the automorphism $\sqrt[g]{qa_2} \leftrightarrow \zeta\sqrt[g]{qa_2}$ we get

$$\begin{aligned} \zeta^i\sqrt[g]{q'a_2} &= a_0 + a_1\zeta\sqrt[g]{qa_2} + \ldots + a_{g-1}\zeta^{g-1}(\sqrt[g]{qa_2})^{g-1} \\ &= \zeta^i(a_0 + a_1\sqrt[g]{qa_2} + \ldots + a_{g-1}(\sqrt[g]{qa_2})^{g-1}). \end{aligned}$$

Because of the independence of 1, $\sqrt[g]{qa_2}, \ldots, (\sqrt[g]{qa_2})^{g-1}$ over $\mathfrak{F}(\zeta)$ we must have

$$\zeta^i a_j = \zeta^j a_j, \; a_j = 0 \text{ for } j \neq i.$$

Hence

$$\begin{aligned} \sqrt[g]{q'a_2} &= a_i(\sqrt[g]{qa_2})^i \\ q'a_2 &= a_i^g(qa_2)^i. \end{aligned}$$

Our choice of q and q', together with equation 14, imply that $i = 1$ and a_i must be a unit. Hence $(q) = (q')$.

Clearly we can choose infinitely many (q) which are square free and prime to a_1, a_2, $\mathfrak{p}$ and the absolute differente of $\mathfrak{F}(\zeta)$. For instance all but a finite number of rational primes fulfill this condition.

The ideal $(a_1, \sqrt[g]{qa_2})$ is moreover a prime ideal since it lies in a field of degree g over $\mathfrak{F}$ and its gth power is a prime ideal in $\mathfrak{F}$. For the same reason it also originates in $\mathfrak{F}$ since it cannot lie in any field of degree less than g over $\mathfrak{F}$.

Theorem 4 shows among other things: If $\mathfrak{p}^h$, $h > 1$, is a prime ideal in $\mathfrak{F}'$ over $\mathfrak{F}$ then $\mathfrak{p}$ originates in infinitely many fields over $\mathfrak{F}$. For let $\mathfrak{p}^g$ be the highest power of $\mathfrak{p}$ which is a prime ideal in some extension of $\mathfrak{F}$. Let $\mathfrak{F}''$ be the unique field over $\mathfrak{F}$ in which $\mathfrak{p}^g$ originates and let $\mathfrak{p}$ originate in some field $\mathfrak{F}_1$ over $\mathfrak{F}''$. By Theorem 4 there are infinitely many such fields. We must show that $\mathfrak{p}$ originates in $\mathfrak{F}_1$ over $\mathfrak{F}$. If $\mathfrak{p}$ lies in $\mathfrak{F}_2$ over $\mathfrak{F}$ where $\mathfrak{F}_1 \supseteq \mathfrak{F}_2$, then $\mathfrak{F}_2 \supseteq \mathfrak{F}''$ by Theorem 2 and hence $\mathfrak{F}_1 = \mathfrak{F}_2$ since $\mathfrak{p}$ originates in $\mathfrak{F}_1$ over $\mathfrak{F}''$. Thus $\mathfrak{p}$ also originates in $\mathfrak{F}_1$ over $\mathfrak{F}$.

Theorem 2 characterizes completely the fields over $\mathfrak{F}$ which contain a given prime ideal $\mathfrak{p}$ if no power of $\mathfrak{p}$ is a prime ideal in a field over $\mathfrak{F}$. However in the case that some $\mathfrak{p}^h$ $(h > 1)$ is a prime ideal in a field over $\mathfrak{F}$ we obtain only the necessary condition that every field containing $\mathfrak{p}$ must contain the field in which $\mathfrak{p}^g$ originates where $\mathfrak{p}^g$ is defined in Theorem 2. A stronger necessary but still not sufficient condition is as follows:

THEOREM 5. *If $\mathfrak{p}$ originates in $\mathfrak{F}'$ over $\mathfrak{F}$, $\mathfrak{p}^g = \mathfrak{P}$ is the highest power of $\mathfrak{p}$ which is a prime ideal in some subfield of $\mathfrak{F}'$ and if $\mathfrak{p}^g$ originates in $\mathfrak{F}''$ then $\mathfrak{F}' = \mathfrak{F}''(\alpha)$, where α satisfies an irreducible equation*

$$x^m + a_1 x^{m-1} + \ldots + a_m = 0 \tag{13}$$

of degree $m = gr$ (r integral) with coefficients in $\mathfrak{F}''$ such that

$$a_{lg+k} \equiv 0(\mathfrak{P}^{l+1}), \quad k > 0,$$
$$a_{rg} \not\equiv 0(\mathfrak{P}^{r+1}). \tag{14}$$

Proof. From Theorem 2 we have $\mathfrak{F}'' \subset \mathfrak{F}'$. Let $\alpha \subset \mathfrak{p}$, α non $\subset \mathfrak{p}^2$, $\alpha \subset \mathfrak{F}'$. Since $\mathfrak{p}$ originates in $\mathfrak{F}'$ and since in every field between $\mathfrak{F}''$ and $\mathfrak{F}'$ the ideal $\mathfrak{p}$ corresponds to a power of $\mathfrak{p}$ we must have $\mathfrak{F}' = \mathfrak{F}''(\alpha)$. Let $(\mathfrak{F}'/\mathfrak{F}'') = m$ and observe that the conjugates of α over $\mathfrak{F}''$ are all exactly divisible by $\mathfrak{p}$. Hence the $(lg + k)$th, $(k > 0)$, symmetric function of these conjugates is divisible by $\mathfrak{p}^{lg+k}$ and since it is in $\mathfrak{F}''$ it must be divisible by $\mathfrak{P}^{l+1}$. Moreover the last coefficient is exactly divisible by $\mathfrak{p}^m$. If $\mathfrak{p} = \mathfrak{p}_1^{e_1} \ldots \mathfrak{p}_s^{e_s}$ is the prime decomposition of $\mathfrak{p}$ in $\mathfrak{F}'$ and f_i the degree of $\mathfrak{p}_i$ then $\mathfrak{p}_i$ is of multiplicity ge_i with respect to $\mathfrak{P}$ and hence

$$m = ge_1 f_1 + \ldots + ge_s f_s = gr \qquad (r \text{ integral}). \tag{15}$$

This proves Theorem 5.

THEOREM 6. *Let $\mathfrak{p}^g = \mathfrak{P}$ and let g and $\mathfrak{F}''$ be defined as in Theorem 5. The ideal $\mathfrak{p}$ lies in $\mathfrak{F}'$ over $\mathfrak{F}$ if and only if $\mathfrak{F}' \supset \alpha$ where $\alpha^g = \beta$ satisfies an irreducible equation*

$$\beta^r + a_1\beta^{r-1} + \ldots + a_r = 0,\ a_i \equiv 0(\mathfrak{P}^i),\ a_r \not\equiv 0(\mathfrak{P}^{r+1}),\ \textit{over } \mathfrak{F}''. \tag{16}$$

First let $\mathfrak{p}$ lie in $\mathfrak{F}'$, then there exists in $\mathfrak{F}'$ an α such that $\alpha \equiv 0(\mathfrak{p})$, $\alpha \not\equiv 0(\mathfrak{p}^2)$. By Theorem 2 we have $\alpha \subset \mathfrak{F}' \subset \mathfrak{F}''$. Clearly $\alpha^g = \beta$ and all its conjugates over $\mathfrak{F}''$ are exactly divisible by $\mathfrak{P}$ and the necessity of the condition 16 follows.

On the other hand consider $\mathfrak{F}''(\alpha)$ where $\alpha^g = \beta$ satisfies an irreducible equation 16. Let γ be a number with ideal denominator $\mathfrak{P}$. Then $\gamma\beta$ satisfies an equation

$$(\gamma\beta)^r + \gamma a_1(\gamma\beta)^{r-1} + \ldots + \gamma^r a_r = 0 \tag{17}$$

with integral coefficients. Hence $\beta \equiv 0(\mathfrak{P})$. Moreover since $a_r \not\equiv 0(\mathfrak{P}^{r+1})$ it follows that $\beta = \mathfrak{P}\mathfrak{b}$, $(\mathfrak{P}, \mathfrak{b}) = 1$. Consider the ideal $(\alpha, \mathfrak{P})$. If

$$\mathfrak{P} = \mathfrak{P}_1^{e_1} \ldots \mathfrak{P}_s^{e_s}$$
$$\alpha = \mathfrak{P}_1^{h_1} \ldots \mathfrak{P}_s^{h_s}\mathfrak{c}, \qquad (\mathfrak{p}_1, \mathfrak{c}) = 1 \tag{18}$$

it follows that $e_i = gh_i$. Hence $(\alpha, \mathfrak{P})^g = \mathfrak{P}$.

Thus $\mathfrak{F}''(\mathfrak{a})$ contains $\mathfrak{p}$ and so does every field over $\mathfrak{F}''(\mathfrak{a})$.

Suppose an ideal $\mathfrak{p}$ a power of which is a prime ideal in some field over $\mathfrak{F}$ is given in any field $\mathfrak{F}_1$ over $\mathfrak{F}$ and we are required to find all extensions of $\mathfrak{F}$ which contain $\mathfrak{p}$. We proceed as follows. We first find the largest power say $\mathfrak{p}^\sigma = \mathfrak{P}$ of $\mathfrak{p}$ which is a prime ideal in any admissible subfield of $\mathfrak{F}_1$. Next we determine the smallest admissible subfield containing $\mathfrak{P}$. Let this field be $\mathfrak{F}''$. We then obtain all fields which contain $\mathfrak{p}$ as all extensions of all $\mathfrak{F}''(\mathfrak{a})$ where $\mathfrak{a}^\sigma$ satisfies an equation of the form 16.

Ohio State University

EXTREMUM PROPERTIES OF THE REGULAR POLYHEDRA

LÁSZLÓ FEJES TÓTH

1. Historical remarks. In this paper we extend some well-known extremum properties of the regular polygons to the regular polyhedra. We start by mentioning some known results in this direction.

First, let us briefly consider the problem which has received the greatest attention among all the extremum problems for polyhedra. It is the determination of the polyhedron of greatest volume V of a class of polyhedra of equal surface areas F, i.e., the isepiphan problem.

The simple fact that the regular tetrahedron is the best among the tetrahedra was already known to Lhuilier.[1] But let us at once note that, among the 8- and 20-cornered polyhedra, the cube and the regular dodecahedron are not the best ones, and similarly, the regular octahedron and icosahedron are not the best polyhedra among the 8- and 20-faced polyhedra.

Steiner,[2] who was certainly in possession of this fact, announced only the conjecture that any regular polyhedron is the best one among the topologically isomorphic polyhedra. In proving this conjecture he succeeded, apart from the tetrahedron, only for the octahedron. The case of the icosahedron is, up to the present day, unsettled.

In 1935, M. Goldberg[3] made an attempt to prove the inequality

$$F^3/V^2 \geqslant 54(f-2)\tan\omega_f(4\sin^2\omega_f - 1);\quad \omega_f = \frac{f}{f-2}\frac{\pi}{6}$$

concerning a convex f-faced polyhedron. This inequality (for which I subsequently gave a complete proof[4]) is exact for $f=4$, 6 and 12 and gives an exact asymptotical estimate for large values of f. Equality holds only for a regular tetrahedron, hexahedron, and dodecahedron.

According to this the regular hexahedron and dodecahedron are proved to be the best not only among the polyhedra of their type but also among all convex polyhedra with 6 and 12 faces, respectively.

Received September 24, 1948. The earlier publications of the author appeared under the name "Fejes". In order to explain this fact the author communicates the following at the request of the editors: Kolozsvár (Roumanian Cluj, the capital of Transsylvania, the native town of J. Bólyai, and where L. Fejér, F. Riesz and A. Haar began their career as young professors) was ceded to Roumania by the Treaty of 1920. From 1940 to 1944 it belonged temporarily to Hungary. The author generally worked in Kolozsvár during the time 1941-1944. Returning to Budapest he took the name "Fejes Tóth" (to be found in old family documents, and already used by some other members of his family), partly in order to avoid confusion with the name of Professor L. Fejér.

[1] S. Lhuilier, *De relatione mutua capacitatis et terminorum figurarum*, etc. (Varsaviae, 1782).

[2] J. Steiner, *Gesammelte Werke* II, 117-308.

[3] M. Goldberg, "The Isoperimetric Problem for Polyhedra," *Tôhoku Math. J.*, vol. 40 (1935), 226-236.

[4] L. Fejes Tóth, "The Isepiphan Problem for n-hedra," *Amer. J. Math.*, vol. 70 (1948), 174-180.

Also, the following inequality

$$F^3/V^2 \geqq \frac{27\sqrt{3}}{2}(v-2)(3\tan^2\omega_v - 1);\ \omega_v = \frac{v}{v-2}\frac{\pi}{6}$$

probably holds for any convex polyhedron with v vertices. It is exact for $v = 4$, 6 and 12 and gives an exact asymptotical estimate for large values of v. This would mean that the regular octahedron and icosahedron are—again far beyond Steiner's conjecture—the best polyhedra among all 6- and 12-cornered polyhedra.

The state of affairs in the isepiphan problem is characteristic of a number of other problems.[5] Therefore in order to give a general orientation in the possibilities of transferring different extremum properties of the regular n-gon to space we can say:

When f is given it is the trihedral-cornered regular polyhedra, and when v is given the triangular-faced, that generally play a prominent part in the solutions of the extremum problems. It is inherent in the problem that—contrary to the problems in the plane—we cannot expect to determine the extremal polyhedra for all values of f or v. We must rather be content with inequalities exact for 4, 6 and 12 and asymptotically exact for large values of f or v.

Let us note that—taking into account the great number of researches dealing with various extremum properties of the regular polygons—it is surprising that, for instance, no extremum property of the regular icosahedron or dodecahedron occurs, as far as I know, in earlier literature. Still less do we find a systematic treatment of such extremum properties. Therefore, much remains to be done in the extremum problems for polyhedra to bring our knowledge, in this respect, to a level with that of the polygons. These attractive questions offer ample scope for work.

2. Aim and results. As we have seen, the researches made hitherto related to polyhedra of a given type or to polyhedra of a given number of faces or vertices.

But the consideration of a type of polyhedra is too special to obtain general results. On the other hand, the class of polyhedra of a given number of faces or vertices is too large to obtain all the five regular polyhedra as solutions of the same extremum problem. Therefore, in the following, we are going to compare polyhedra having a given number of faces f *and* a given number of vertices v. In this way we shall obtain inequalities in which equality holds for *all the five regular solids*.

In this paper we shall prove the following

THEOREM. *If V denotes the volume, r the radius of the insphere and R the radius of the circumsphere of a convex polyhedron having f faces, v vertices and e edges, then*

[5] See, for instance, the paper L. Fejes Tóth, "An Inequality Concerning Polyhedra," *Bull. Amer. Math. Soc.*, vol. 54 (1948), 139-146, where further bibliographical data can be found.

$$V \geqslant \frac{e}{3} \sin \frac{\pi f}{e} \left(\tan^2 \frac{\pi f}{2e} \tan^2 \frac{\pi v}{2e} - 1 \right) r^3 \tag{1}$$

$$V \leqslant \frac{2e}{3} \cos^2 \frac{\pi f}{2e} \cot \frac{\pi v}{2e} \left(1 - \cot^2 \frac{\pi f}{2e} \cot^2 \frac{\pi v}{2e} \right) R^3. \tag{2}$$

Equality holds in both inequalities only for the regular polyhedra.

Letting $p = 2e/f$ and $q = 2e/v$, we obtain by combining (1) and (2) the following

COROLLARY. *If r and R denote the radii of the in- and circumsphere of a convex polyhedron for which the average number of the sides of the faces and the average number of the edges of the vertices is p and q, respectively, then*[6]

$$\frac{R}{r} \geqslant \tan \frac{\pi}{p} \tan \frac{\pi}{q}. \tag{3}$$

Professor H. S. M. Coxeter wrote to me calling my attention to the equality $R/r = \tan \pi/p \tan \pi/q$ which holds for any regular polyhedron having p-gonal faces, q at each vertex. By this remark I was impelled to prove the nice inequality (3) which was the point of departure of the present paper.

3. Proofs. In order to prove (1) we may obviously suppose—without loss of generality—that the insphere of centre O has the radius $r = 1$. Denote the faces of the polyhedron Π, and their area as well by F_i $(i = 1, 2, \ldots, f)$, the solid angle under which F_i appears from O by σ_i, and the number of the sides of the polygon F_i by p_i.

It is easy to see that for given values of p_i and σ_i the area F_i takes its minimum if F_i is a regular p_i-gon touching the insphere at its own centre. This minimum property is expressed—as a simple computation shows—by the inequality

$$F_i \geqslant \Phi(\sigma_i, p_i); \quad \Phi(\sigma, p) = \frac{p}{2} \sin \frac{2\pi}{p} \left(\tan^2 \frac{\pi}{p} \cot^2 \frac{2\pi - \sigma}{2p} - 1 \right).$$

Now we make use of the fact that the function of two variables $\Phi(\sigma, p)$ is convex from below for $0 \leqslant \sigma \leqslant 2\pi$, $3 \leqslant p$. Hence by Jensen's inequality[7]

$$3V \geqslant \sum_{i=1}^{f} F_i \geqslant \sum_{i=1}^{f} \Phi(\sigma_i, p_i) \geqslant f \Phi(4\pi/f, 2e/f); \qquad \text{q.e.d.}$$

The only difficulty of this very simple proof—which is properly Goldberg's proof mentioned above—is the unfortunate circumstance that the function

[6] The inequality (3) is a generalization of the inequality $R \geqslant 3r$ concerning tetrahedra—found in 1943 by a young Hungarian mathematician I. Ádám at the suggestion of Professor L. Fejér—and of certain results of the author (see the paper referred to in footnote 5).

[7] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Math.*, vol. 30 (1906), 175-193.

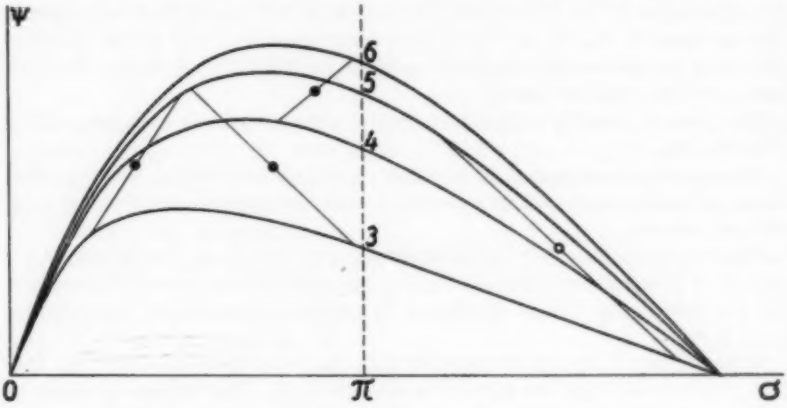
$\Phi(\sigma, p)$ is too complicated to arrange clearly the computations necessary for the proof of convexity.[8] On the other hand, it is easy to give a graphical representation of $\Phi(\sigma, p)$ from which the convexity can be seen empirically.[9]

Let us now turn to the inequality (2), the proof of which is analogous to the foregoing. Decompose Π into f pyramids of volumes $V_1, V_2, \ldots V_f$, having the centre O of the circumsphere of radius $R = 1$ as a common vertex, with bases formed by the respective faces $F_1, F_2, \ldots F_f$ of Π.

We have now the inequality

$$V_i \leqslant \Psi(\sigma_i, p_i); \quad \Psi(\sigma, p) = \frac{p}{3} \cos^3 \frac{\pi}{p} \tan \frac{2\pi - \sigma}{2p} \left(1 - \cot^2 \frac{\pi}{p} \tan^2 \frac{2\pi - \sigma}{2p}\right)$$

which means that, for given values of p_i and for given values of the area σ_i of the projection of F_i from O upon the circumsphere, the volume V_i takes its maximum if F_i is a regular p_i-gon the vertices of which lie on the circumsphere.



But now in addition to the difficulty indicated in the above proof a further one arises, namely: the function $\Psi(\sigma, p)$, as a function of two variables, is not convex from above in the whole strip $0 \leqslant \sigma \leqslant 2\pi$, $p \geqslant 3$. But it will be sufficient to make use of the convexity, say, for $0 \leqslant \sigma \leqslant \pi$, which can be surmised with great confidence from the above graphical representation of a few functions $\Psi(\sigma, \text{const.})$. The convexity is expressed by the fact that, for instance, the midpoint of any segment joining a point of the curve $\Psi = \Psi(\sigma, p_1)$ with a point of $\Psi = \Psi(\sigma, p_2)$ lies always below the curve $\Psi = \Psi\left(\sigma, \frac{p_1 + p_2}{2}\right)$.

[8] On this occasion I take the liberty to cite from the letter of M. Goldberg written to me in connection with my paper referred to in footnote 4: "Your rigorous proof ... has removed a difficulty which I have tried to overcome without success."

[9] See my paper: "Über einige Extremaleigenschaften der regulären Polyeder und des gleichseitigen Dreiecksgitters," *Annali della Scuola Norm. Sup. di Pisa* (2) 13 (1948), 51-58.

First of all, we are going to prove the inequality (2) for $f \geqslant 8$. Let us note for this purpose that for any value of $p \geqslant 3$ we have the inequalities

$$\Psi(\sigma, p) \leqslant \Psi(\pi, p) \qquad \text{for} \quad \sigma \geqslant \pi,$$
$$\Psi(\sigma_1, p) \leqslant \Psi(\sigma_2, p) \qquad \text{for} \quad 0 \leqslant \sigma_1 \leqslant \sigma_2 \leqslant \pi/2.$$

Let us replace any value $\sigma_i > \pi$ by π. Let us denote the new values by $\sigma'_1, \sigma'_2, \ldots, \sigma'_f$ and their sum by $\sigma' (\leqslant 4\pi)$. Owing to the above inequalities, we have for $f \geqslant 8$

$$V = \sum_{i=1}^{f} V_i \leqslant \sum_{i=1}^{f} \Psi(\sigma_i, p_i) \leqslant \sum_{i=1}^{f} \Psi(\sigma'_i, p_i) \leqslant f\Psi(\sigma'/f, p) \leqslant f\Psi(4\pi/f, p).$$

This is just the inequality (2).

The detailed discussion of the several types of polyhedra for which $f < 8$ contains no interesting new ideas. Instead of such a discussion let us consider, for example, only the type of a 5-sided prism ($f = 7$, $v = 10$), or more generally the case $f \geqslant 6$, $p \geqslant 4$. Since, for a fixed value of p ($p \geqslant 4$), the function $\Psi(\sigma, p)$ is an increasing function of σ up to a constant $c_p \geqslant 2\pi/3$, the proof runs word for word as above.

The cases of equality are evident, by the above proofs, in both inequalities (1) and (2).

Now we are going to give two further rigorous proofs of (1). On the other hand, we must admit that an attempt at a similar proof of the inequality (2) did not succeed.

Again let O be the centre of the insphere and put $r = 1$. Let us consider a face F_i of Π and denote the foot of the perpendicular from O to the face-plane by A. Further, let BD be an edge of F_i and C the foot of the perpendicular from A on it.

Suppose that C lies on the segment BD, just as A lies within F_i, and that this proves to be right for all faces and edges of Π. The surface of Π can in this case be decomposed into $4e$ right triangles one of which is ABC.

Consider the right spherical triangle $A'B'C'$ arising by central projection of ABC from O upon the insphere. Denote the angle at A' by α, the angle at B' by β and the hypotenuse $A'B'$ by c. Since $AB \geqslant \tan c$ and $\cos c = \cot\alpha \cot\beta$, the area t of the triangle ABC is given by

$$t \geqslant \tfrac{1}{4} \sin 2\alpha \tan^2 c = \tfrac{1}{4} \sin 2\alpha \, (\tan^2\alpha \tan^2\beta - 1) \equiv \Theta(\alpha, \beta).$$

Furthermore, since

$$\Theta_{\alpha\alpha}\Theta_{\beta\beta} - \Theta_{\alpha\beta}^2 = \frac{2\tan^4\alpha}{\cos^6\beta} [1 - (\sin^2\alpha + \sin^2\beta)]^2 \geqslant 0,$$

the function $\Theta(\alpha, \beta)$ in the domain determined by the inequalities $0 \leqslant \alpha \leqslant \pi/2$, $0 \leqslant \beta < \pi/2$, $\alpha + \beta \geqslant \pi/2$, is convex from below[10] and we have

[10] For the transformation of the Jacobian $\Theta_{\alpha\alpha}\Theta_{\beta\beta} - \Theta_{\alpha\beta}^2$ into the above simple form I am obliged to Mr. J. Molnár.

$$3V \geqq \sum t \geqq 4e\Theta(2\pi f/4e, 2\pi v/4e).$$

This is just the inequality (1) to be proved.

In order to get rid of the above restriction concerning the feet of the perpendiculars we can use the inequalities $F_i \geqq \Phi(\sigma_i, p_i)$ of the first proof. In other words, we can replace any face by an admissible polygon of smaller area of which the number of sides and the area of the projection remain invariant.

The following alternative proof makes no use of the discussion of any special function.[11] We shall obtain the inequality in question as a corollary of the following general

THEOREM. *Decompose the surface S of the unit sphere by a net N having v vertices and e edges into a finite number $f \geqq 4$ of convex spherical polygons $\sigma_1, \sigma_2, \ldots, \sigma_f$. Further let $P_1, P_2, \ldots, P_f$ be f points of S and $\varphi(\rho)$ a strictly increasing function defined for $0 \leqq \rho < \pi$. Then*

$$\sum_{i=1}^{f} \int_{\sigma_i} \varphi(P_iP)d\omega \geqq 4e \int_{\Delta} \varphi(AP)d\omega \tag{4}$$

where $d\omega$ denotes the area element of S at the variable point P, and Δ a right spherical triangle ABC the acute angles of which are $\alpha = \pi f/2e$ at A and $\beta = \pi v/2e$ at B. Equality holds only if N is the central projection of the edges of a regular polyhedron circumscribed about S and $P_1, P_2, \ldots, P_f$ the points of contact of the faces of this polyhedron.

Preparatory to the proof we make two remarks, easy to prove, in which a spherical domain and its area are denoted by the same symbol.

REMARK 1. Let s be a segment of a spherical cap c ($< 2\pi$) with the top point T. Then the function

$$\Omega(s) = \int_s \varphi(TP)d\omega$$

is convex from above for $0 \leqq s \leqq c/2$.

REMARK 2. For any convex domain d lying in a "hemicap" of c,

$$\int_d \varphi(TP)d\omega \leqq \Omega(d).$$

Let us first note that the integral $\int_{\sigma_i} \varphi(P_iP)d\omega$ obviously takes its minimum for a variable P_i at an inner point of σ_i. Therefore we may suppose that P_i lies within σ_i ($i = 1, 2, \ldots, f$).

Let c_i be the spherical cap with the top point P_i and the radius AB, while $Q_1, Q_2, \ldots Q_{p_i}$ are the vertices of σ_i and $s_1, s_2, \ldots, s_{p_i}$ are the convex partial-domains of c_i lying outside of σ_i, the first bordered by the great circles P_iQ_1, Q_1Q_2, P_iQ_2, the second by P_iQ_2, Q_2Q_3, P_iQ_3, etc. Omitting the common integrand $\varphi(P_iP)d\omega$ under the integral signs we have

[11] Cf. the proof in the paper referred to in footnote 4.

$$\int_{\sigma_i} = \int_{c_i} - \sum_{\nu=1}^{p_i} \int_{s_\nu} + \int_{\sigma'_i}$$

where σ'_i denotes the part of σ_i not covered by c_i.

Consider the corresponding equalities for $i = 1, 2, \ldots, f$. The total number of the domains s_ν being $2e$, we get by the above remarks and Jensen's inequality

$$\sum_{i=1}^{f} \int_{\sigma_i} = f \int_c - \sum_{\nu=1}^{2e} \int_{s_\nu} + \sum_{i=1}^{f} \int_{\sigma'_i} \geqslant f \int_c - \sum_{\nu=1}^{2e} \Omega(s_\nu) + \sum_{i=1}^{f} \int_{\sigma'_i}$$

$$\geqslant f \int_c - 2e\,\Omega\left(\sum_{\nu=1}^{2e} \frac{s_\nu}{2e}\right) + \sum_{i=1}^{f} \int_{\sigma'_i},$$

where c denotes a spherical cap of radius AB and top point A, and

$$\int_c = \int_c \varphi(AP)d\omega.$$

Since, with the notation $\sum_{i=1}^{f} \sigma'_i = S'$, we have

$$S = fc - \sum_{\nu=1}^{2e} s_\nu + S',$$

we may write

$$\sum_{i=1}^{f} \int_{\sigma_i} \geqslant f \int_c - 2e\Omega\left(\frac{fc - S + S'}{2e}\right) + \sum_{i=1}^{f} \int_{\sigma'_i}$$

$$= f \int_c - 2e\Omega\left(\frac{fc - S}{2e}\right) + \sum_{i=1}^{f} \int_{\sigma'_i} - 2e \int_d \varphi(AP)d\omega,$$

denoting by d the partial domain of c which completes the segment of the cap c of area $(fc - S)/2e$ to the segment of area $(fc - S + S')/2e$. Furthermore, since by the monotonicity of $\varphi(\rho)$ the sum of the last two terms in the above inequality is $\geqslant 0$, we have

$$\sum_{i=1}^{f} \int_{\sigma_i} \geqslant f \int_c - 2e\,\Omega\left(\frac{fc - S}{2e}\right) = 4e\left\{\frac{\alpha}{2\pi} \int_c - \tfrac{1}{2}\Omega\left(\frac{fc - S}{2e}\right)\right\}.$$

But $(fc - S)/2e$ equals the area of the segment s of the cap c cut off by BC. This is obvious by

$$\Delta = \alpha + \beta - \frac{\pi}{2} = \frac{\alpha}{2\pi} c - \tfrac{1}{2} s.$$

This completes the proof of (4).

Equality holds only if S is entirely covered by the caps c_i without any part being covered three times, and the domains s_ν are all congruent segments of a spherical cap. This is just the case indicated above.

The inequality (1) is an immediate consequence of (4) for the function $\varphi(\rho) = \sec^3\rho$, for which $\frac{1}{3}\int_d \varphi(TP)d\omega$ equals the volume of the cone with a

vertex at the centre of S cutting out from S the domain d and the base plane of which touches S at the point T.

Let us still note that the consideration of the function

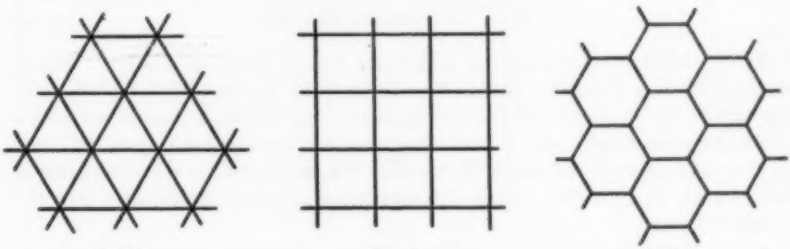
$$\varphi(\rho) = \begin{cases} \sec^3 \rho & \text{for } 0 \leqslant \rho \leqslant AB \\ \sec^3 AB & \text{for } AB \leqslant \rho \leqslant \pi/2 \end{cases}$$

involves a sharpening of (1), according to which the volume V of Π can be replaced by the volume of the part of Π which lies in a sphere of radius $r \tan \frac{\pi}{p} \tan \frac{\pi}{q}$ concentric with the insphere.

4. The regular degenerate polyhedra.[12] The five Platonic solids can be supplemented in a natural manner by three further "polyhedra" inscribed in or circumscribed to the sphere of infinite radius, or more correctly: tessellations in the plane. If we denote by $\{p, q\}$ the regular polyhedron having p-gonal faces, q at each vertex, then the eight regular polyhedra can be arranged into the following scheme:

$$\begin{array}{cccc} \{3,3\} & \{3,4\} & \{3,5\} & \{3,6\} \\ \{4,3\} & \{4,4\} & & \\ \{5,3\} & & & \\ \{6,3\} & & & \end{array}$$

Any of the three regular degenerate polyhedra, represented by the figures below, can be considered as the limiting form of a set of convex polyhedra.



$\{3,6\}$ $\{4,4\}$ $\{6,3\}$

The introduction of this terminology will prove suitable in the investigation of the question when our inequalities give exact asymptotic estimates for large values of e. Let us consider, for instance, a set of polyhedra of increasing values of e for which

$$\lim e \left(\frac{R}{r} - \tan \frac{\pi}{p} \tan \frac{\pi}{q} \right) = 0.$$

[12] Cf. H. S. M. Coxeter, *Regular Polytopes* (New York, 1949), chap. IV.

The "limiting poryhedra" of such sets are the regular ones.

In a certain sense we can briefly say that equality holds in our inequalities for, and only for, the eight regular polyhedra.

5. Further problems. Let us consider the inequalities analogous to (1) and (2) for the surface area F of the polyhedron Π:

$$e \sin\frac{2\pi}{p}\left(\tan^2\frac{\pi}{p}\tan^2\frac{\pi}{q}-1\right)r^2 \leqq F \leqq e\sin\frac{2\pi}{p}\left(1-\cot^2\frac{\pi}{p}\cot^2\frac{\pi}{q}\right)R^2.$$

The inequality on the left is equivalent to (1). On the other hand, the inequality on the right seems to involve some difficulties. We are going to prove this inequality only for polyhedra the faces and edges of which contain the foot of the centre of the circumsphere on their plane or line, respectively.

The proof is a dual counterpart of the second proof of (1). Let us keep the notations of this proof surrendering the rôle of the insphere to the circumsphere of radius $R = 1$. We have now $AB \leqq \sin c$ and hence

$$t \leqq \tfrac{1}{4}\sin 2\alpha \sin^2 c = \tfrac{1}{4}\sin 2\alpha\,(1-\cot^2\alpha\cot^2\beta)$$
$$= \Gamma(\alpha,\beta) = -\Theta\left(\frac{\pi}{2}-\alpha,\frac{\pi}{2}-\beta\right).$$

Since $\Gamma_{\alpha\alpha}\Gamma_{\beta\beta}-\Gamma_{\alpha\beta}^2 = \dfrac{2\cot^4\alpha}{\sin^6\beta}[1-(\cos^2\alpha+\cos^2\beta)]^2 \geqslant 0$, the function $\Gamma(\alpha,\beta)$, for $0 \leqq \alpha < \pi/2,\ 0 \leqq \beta < \pi/2,\ \alpha+\beta \geqslant \pi/2$, is concave from below and we have

$$F = \sum t \leqq 4e\,\Gamma(2\pi f/4e, 2\pi v/4e); \qquad \text{q.e.d.}$$

The proof of the general case miscarries for the following reason. Let us change the face F_i within the insphere so that the number p_i of its sides and the area σ_i of its projection from O upon the circumsphere remain invariant. Then the area F_i has only a local maximum for the regular p_i-gon inscribed in the circumsphere and takes its absolute maximum just in the case when A lies outside F_i, provided that σ_i remains below a certain constant which depends only on p_i.

Let us now return to the isepiphan problem. According to a well-known result of L. Lindelöf[13] the f-hedron which minimizes, by a given value of f, the quotient F^3/V^2 has the property of being circumscribed about a sphere. Hence for the best f-hedron $F^3/V^2 = 9F/\pi^2$. But this holds not only for the best f-hedra, but also, for instance, for the best dipyramids of given number of vertices and for the best polyhedra among many other classes of polyhedra as well. All these induce us to announce the following conjecture concerning any convex polyhedron:

[13] L. Lindelöf, "Propriétés générales des polyèdres etc.," *St. Petersburg Bull. Acad. Sci.*, vol. 14 (1869), 258-269.

$$F^3/V^2 \geqslant 9e \sin\frac{2\pi}{p}\left(\tan^2\frac{\pi}{p}\tan^2\frac{\pi}{q} - 1\right).$$

The proof of this inequality would, in a certain sense, close the range of the isepiphan problems for polyhedra.

Let us now agree upon the notations $A(x; k)$ and $H(x; k)$ for the arithmetic and harmonic means of certain numbers x_i with the weights k_i. Let further Π be a convex polyhedron, O an arbitrary inner point of it, $p_1, p_2, \ldots, p_f$ the numbers of the sides of the faces distant $r_1, r_2, \ldots, r_f$ from O, $q_1, q_2, \ldots, q_v$ the numbers of the edges running into the vertices distant $R_1, R_2, \ldots, R_v$ from O and put, as above, $p = A(p; 1)$, $q = A(q; 1)$. With these notations the following inequality probably holds:

$$A(R; q)/H(r; p) \geqslant \tan\frac{\pi}{p}\tan\frac{\pi}{q}.$$

This may be a generalization of certain previous results[14] suggested by a triangle inequality of L. J. Mordell and P. Erdös.[15] Here $A(R; q)$ cannot be replaced by $A(R; 1)$ just as $H(r; p)$ cannot be replaced by $H(r; 1)$. Similarly, the above inequality cannot be improved by putting $H(R; q)$ instead of $A(R; q)$ or $A(r; p)$ instead of $H(r; p)$.

[14] L. Fejes Tóth, "Inequalities Concerning Polygons and Polyhedra," *Duke Math. J.*, vol. 15 (1948), 817-822.

[15] L. J. Mordell, Problem 3740, proposed by Paul Erdös, *Amer. Math. Monthly*, vol. 44 (1937), 252.

Budapest

ON FREQUENCIES AND SEMICONTINUOUS FUNCTIONS

F. W. LEVI

THIS paper deals with a particular class of distributive[1] properties which appear to be important for Analysis and which I call *frequencies*. They can be defined for any kind of sets but it essential for proper application that a condition **L** (the statement of Lindelöf's lemma)[2] is satisfied. From this condition follows Theorem 1, which is characteristic for frequencies but does not hold for other distributive properties. For every frequency F of a space $\sum$, one can build up an Analysis mod F of $\sum$; the classical case is the Analysis mod F_0. It is convenient to introduce the words "nearly every" with such a meaning that "every" and "almost every" are the special cases which, when we use the notation of this paper, correspond to $F = F_0$ and $F = F_C$. These notations are applied to the semicontinuous functions which are obtained by the upper and lower limiting operations and their iteration. In this way an appropriate tool for investigating the discontinuities of a function is obtained. The iteration of the limiting process leads to interesting "pairs" of functions which are the upper (lower) limiting functions of a set of functions. The coordination into pairs is independent of the frequency F, a fact which proves to be important for the investigation of the pairs. The notion of frequency is also useful for other purposes, e.g. for a generalization of uniform convergence.

1. Consider a set $\sum$ (called space) of elements (called points) in which a family of subsets (called open sets) is distinguished which satisfy the following condition:

L. If A is the join of an aggregate of open sets O_ν, then there exists a countable subset of sets O_ν such that A is the join of them.

This condition is satisfied e.g. for locally compact metric spaces when "open" has the usual meaning (*Lindelöf's lemma*).

We use in this paper the symbols $\cap$ and $\cup$ for the set-theoretical "meet" and "join." In particular $\cup_n A_n$ will denote the join of a countable aggregate of sets $A_1, A_2, \ldots$.

A property which, for a subset of $\sum$, either holds or does not hold, is called a *frequency* F when it satisfies the following three conditions:

Condition 1. F holds in $A = \cup_n A_n$ if and only if F holds in at least one A_n.

Condition 2. F does not hold in the empty set.

Condition 3. F holds in $\sum$.

Received September 29, 1948.

[1] Regarding distributive properties, see [2] pp. 9-14 and the literature quoted on p. 48.

[2] See [1] p. 46, [2] p. 38.

Conditions 2 and 3 are introduced for convenience only to exclude properties which either hold for every subset of $\sum$ or for none. Every frequency is a distributive property, but the converse does not hold. If F holds in $S \subseteq A$, it holds also in $A = S \cup A$. The property "to be infinite" which plays an important role in Analysis, is distributive, but not a frequency. Note the following frequencies which may hold in $\sum$:

F_0 = the property "not to be empty",

F_1 = the property "not to be countable",

and—in a space admitting a regular measure function (Caratheodory)—

F_C = the property "to have a positive measure".

If S has the frequency F, then the property of A that $A \cap S$ has the frequency F, is a frequency $F(S)$. If A has the frequency $F(S)$ we say also: "S has the frequency F in A". If every open set which contains a point $x \in \sum$ has the frequency $F(S)$, we say: "S has the frequency F *at* x". That F_a *implies* F_b, is denoted by $F_a \subseteq F_b$; in particular,

$$F_a \subseteq F_0. \tag{1}$$

The frequencies form a partially ordered set, which is not a lattice. The property that a set has two given frequencies F_a and F_b, is not necessarily a frequency; however the property that it has F_a *or* F_b, is a frequency:

$$F_a \cup F_b \supseteq F_a. \tag{2}$$

If $S = S_1 \cup S_2$ has the frequency F, but S_2 has not the frequency F, then $F(S) = F(S_1)$.

If there exists a countable set T in which F holds, then F holds also for some one-point-set $\{x\}$, where $x \in T$. Denote the join of the one-point-sets which have the frequency F, by S_0. If $A \subseteq \sum - S_0$ has the frequency F, then A is non-countable. If $B \cap S_0$ is non-empty, then B has the frequency F. Thus if S_0 is non-empty,

$$F = F_0(S_0) \cup F(\sum - S_0), \text{ where } F(\sum - S_0) \subseteq F_1(\sum - S_0). \tag{3}$$

As **L** holds in $\sum$, the join of the open sets O_n which have not the frequency F, can be represented by

$$\sum'' = \cup_n O_n,$$

and from condition 1 it follows that $\sum''$ has not the frequency F. Thus for $\sum' = \sum - \sum''$, $F = F(\sum) = F(\sum')$. Moreover, $x \in \sum'$ if and only if $\sum$ has the frequency F at x. Hence:

THEOREM 1. *The points at which the space* $\sum$ *has the frequency* F, *form a subset* $\sum'$, *such that* $\sum'$ *has the frequency* F *at each of its points.* $\sum'$ *is non-empty.*

By applying Theorem 1 to the frequency $F(S)$, we obtain the following corollary:

COROLLARY. If S has the frequency F, then the points of S at which S has the frequency F form a non-empty subset S' and S' has the frequency F at each of its points.

Theorem 1 is well known for topological spaces admitting **L** when $F = F_1(B)$. In this case $\sum'$ is the set of the points of condensation of B; moreover it is known for $F = F_C(B)$. That the theorem cannot be generalized to distributive properties which are not frequencies, is seen from the property "to contain an infinite subset of B". This property holds at every limiting point of B, but these may be finite in number.

Given a frequency F, the word *nearly* will mean: *except a set which has not the frequency F*. Thus for $F = F_C$, "nearly every" becomes synonymous with "almost every", whereas for $F = F_0$, it means "every".

2. The frequencies are closely connected with the theory of measure; they can even be considered as special cases of a generalized measure theory which includes both the measure (Lebesgue, Haar etc.) in a locally compact metric space as well as the frequencies.

A measure function μ^* is a set function which for every $\mathfrak{a} \subseteq \sum$ takes only real non-negative numbers and $+\infty$ as values. It is supposed to satisfy the conditions:

I $\mu^*(\mathfrak{a}) \neq 0$ for a suitable $\mathfrak{a}$,

II $\mu^*(\mathfrak{a}) \leqq \mu^*(\mathfrak{a} \cup \beta)$,

III $\mu^*(\bigcup_n \mathfrak{a}_n) \leqq \sum \mu^*(\mu_n)$,

IV If ω is open and $\mathfrak{a} \subseteq \omega$, $\beta \subseteq \sum - \omega$ are compact, then $\mu^*(\mathfrak{a} \cup \beta) = \mu^*(\mathfrak{a}) + \mu^*(\beta)$.

The theory can be generalized; the values of $\mu^*(\mathfrak{a})$ may belong to any linearly ordered system V provided an *addition* for every countable subset of V is defined and this addition satisfies the condition:

$$\sum_n a_n \geqq a_n. \tag{4}$$

V may consist of two elements only, say 0 and 1, and a sum may be equal to 1 if and only if at least one of the terms is equal to 1; e.g., we may put $\mu^*(\mathfrak{a}) = 1$ when $\mathfrak{a}$ has the frequency F, otherwise 0. The theory of measurable sets can be developed without any reference to properties of real numbers other than those supposed to hold for V. From any generalized measure function we can deduce frequencies in the following way. We subdivide V into a well-ordered sequence of "sections", say

$$\ldots \subset S_\nu \subset S_{\nu+1} \subset \ldots,$$

such that every section is closed for addition and contains, with every element a, also the elements $\leqq a$. The property $\mu^*(\mathfrak{a}) > S_k$ is a frequency $F^{(k)}$ with $F^{(k)} \supseteq F^{(\lambda)}$ for $K \leqq \lambda$. The values used in the classical theory form only two such sections and therefore give rise to a single frequency ($\mu^*(\mathfrak{a}) > 0$).

The theory of frequencies admits a modification when we restrict the notion of "subset of $\sum$" to that of "*admissible* subset of $\sum$". Every family ϕ of subsets may be taken as admissible when:

(a) Every open set belongs to ϕ,

(b) If $A_1, A_2, \ldots$ belong to ϕ, then $\cup_n A_n$ belongs to ϕ.

A similar restriction has been applied successfully in the theory of the distributive properties.

3. Let $\sum$ be a topological space which satisfies the second axiom of countability; then **L** holds for open sets O_n. Every open set which contains a point x, will be called a *neighbourhood* of x. Let $C \subseteq \sum$ have the frequency F at each of its points. We consider functions $f(x), g(x), \ldots$ whose domain is C and whose values are real numbers, $+\infty$ or $-\infty$.

We define $f_1(x)$ as the *upper limiting function* mod F of $f(x)$ and $f_2(x)$ *as the lower limiting function* mod F of $f(x)$ in the following way:

$f_1(x)$ is the l.u.b. of the values k which satisfy the condition that for every positive ϵ, the points x' for which

$$f(x') > k - \epsilon \tag{5}$$

have the frequency F at x; $f_2(x)$ is the g.l.b. of the values g which satisfy the condition that for every positive ϵ, the points x'' for which

$$f(x'') < g + \epsilon, \tag{5'}$$

have the frequency F at x. The (upper and lower) limiting functions of the classical theory[3] are those mod F_0. If $f_1(x_0) < h$, then the set of points x, for which $f(x) \geqslant h$, does not have the frequency F at x_0; therefore the set of points x'', for which $f(x'') < h$, has the frequency F, hence $f_2(x_0) < h$. As this holds for every $h > f_1(x_0)$, $x_0 \in A$, it follows that

$$f_2(x) \leqslant f_1(x), x \in C. \tag{6}$$

It is convenient to use the notation

$$g(x_0) \prec f(x_0) \tag{7}$$

when there exists a neighbourhood Ω of x_0 such that $g(x) \leqslant f(x)$ for nearly every $x \in \Omega$. The relation (7) is therefore not a relation between the values $g(x_0)$ and $f(x_0)$, but between the pairs $\{g,x_0\}$ and $\{f,x_0\}$. If (7) and $f(x_0) \prec g(x_0)$ hold, we write

$$g(x_0) \sim f(x_0); \tag{7'}$$

in this case there exists a neighbourhood Ω' of x_0 such that $g(x) = f(x)$ for nearly every $x \in \Omega'$; (7) implies that $g(x) < f(x)$ for every $x \in \Omega$ and (7') implies that $g(x) \sim f(x)$ for every $x \subset \Omega'$. Moreover it follows from the definition of the upper and lower limiting functions mod F that (7) implies

[3] See [1], p. 122.

$$g_1(x) \leqslant f_1(x),\ g_2(x) \leqslant f_2(x),\ x \in \Omega, \tag{8}$$

and that (7′) implies

$$g_1(x) = f_1(x),\ g_2(x) = f_2(x),\ x \in \Omega'. \tag{8'}$$

THEOREM 2. *For nearly every $x \in C$, we have $f(x) \leqslant f_1(x)$, $f_2(x) \leqslant f(x)$, and therefore $f_2(x) < f(x) < f_1(x)$ for $x \in C$.*

Proof. By symmetry it suffices to prove the first statement of the theorem. If for some x, $f(x) > f_1(x)$, then $f(x) \neq -\infty$, $f_1(x) \neq +\infty$, and we can therefore represent C as the join of the (disjoint) sets:

$$C = \cup_n C_n \cup C_\infty \cup C^* \cup C_0,$$

where $x \in C^*$ if and only if $f_1(x) < f(x) = +\infty$,

$x \in C_\infty$ " " " " $-\infty = f_1(x) < f(x)$,

$x \in C_n$ " " " " $\dfrac{1}{n-1} > f(x) - f_1(x) \geqslant \dfrac{1}{n}$ for $n = 1, 2, \ldots$,

$x \in C_0$ " " " " $f_1(x) \geqslant f(x)$.

We prove that none of these sets has the frequency F except C_0. Suppose C^* has the frequency F; then it follows from Theorem 1 that there exists an $x_0 \in C^*$ at which C^* has the frequency F, but as $f(x) = +\infty$ for $x \in C^*$, $f_1(x_0) = +\infty$, contrary to the supposition. If C_∞ or C_n has the frequency F, we partition the set $C_\infty = \cup_m C_{\infty,m}$, $C_n = \cup_m C_{n,m}$ where the second index indicates that $m - 1 < f(x) \leqslant m$ holds ($m = 0, +1, +2, \ldots$). If C_∞ has the frequency F, then some $C_{\infty,m}$ has the frequency F and therefore there exists some $y \in C_{\infty,m}$ at which $C_{\infty,m}$ has the frequency F, but then $f_1(y) \geqslant m - 1 \neq -\infty$. Suppose now that $C_{n,m}$ has the frequency F and $z \in C_{n,m}$; then $f_1(z) \leqslant f(z) - 1/n$. Therefore, if U is any neighbourhood of z, the points $z_1 \in C' = U \cap C_{n,m}$ for which $f(z_1) \leqslant f(z) - 1/n$, have the frequency F. In every neighbourhood U' of z_1 again, the points $z_2 \in C'' = U' \cap C'$, for which $f(z_2) \leqslant f(z_1) - 1/n \leqslant f(z) - 2/n$, have the frequency F; after n steps we find a subset $C^{(n)} \subseteq C_{n,m}$ in which the points $z_n \in C^{(n)}, f(z_n) \leqslant f(z) - 1$ have the frequency F. Hence $C_{n,m} \cap C_{n,m+1}$ is non-empty, contrary to the definition of $C_{n,m}$. Therefore $C_{n,m}$ has not the frequency F. Thus $C - C_0$ has not the frequency F.

As $f(x) < f_1(x)$, it follows from (8) that $f_1(x) \leqslant f_{11}(x)$ throughout C. Suppose that for some $x \in C$, $f_1(x) = c$, $f_{11}(x) = c + 3k$, $k > 0$; then in every neighbourhood Ω of x, the set of points x' for which $f_1(x') \geqslant c + 2k$ has the frequency F, and therefore in every neighbourhood $\Omega' \subseteq \Omega$ of x', the set of points x'' for which $f(x'') > c + k$ has the frequency F. Thus the set of these points x'' has the frequency F at x and therefore $f_1(x) > c$, contrary to the supposition. Hence:

$$f_{11}(x) = f_1(x),\ f_{22}(x) = f_2(x) \tag{9}$$

for every $x \in C$. If now $g_1(x_0) < f_1(x_0)$, then it follows from (8) that $g_{11}(x) \leqslant f_{11}(x)$, $x \in \Omega$ (neighbourhood of x_0), and from (9) that $g_1(x) \leqslant f_1(x)$. Hence:

THEOREM 3. *$g_i(x) < f_i(x_0)$ implies that there exists a neighbourhood Ω of x_0, such that $g_i(x) \leqslant f_i(x)$, $x \in \Omega$, $i = 1$ or 2.*

When we consider several frequencies F_a, F_b, . . . which C has at every $x \in C$, the limiting functions will be distinguished by an upper index. Suppose $F_a \subseteq F_b$; the set of points x'' for which $f(x'') < f^a{}_2(x) + \epsilon$, $\epsilon > 0$ has the frequency F_a and therefore also the frequency F_b at $x \in C$. Therefore $f^b{}_2(x) \leqslant f^a{}_2(x)$.

Hence

$$F_a \subseteq F_b \text{ implies } f^b{}_2(x) \leqslant f^a{}_2(x) \leqslant f^a{}_1(x) \leqslant f^b{}_1(x),\ x \in C. \tag{10}$$

THEOREM 4. *$F_a \subseteq F_b$ implies $f^a{}_i(x) = (f^a{}_i)^b{}_i(x)$, for $x \in C$, $i = 1$ or 2.*

Proof. It suffices to consider $i = 1$. From (9) and (10) we deduce

$$f^a{}_1(x) = (f^a{}_1)^a{}_1(x) \leqslant (f^a{}_1)^b{}_1(x).$$

As in the proof of (9), we put $f^a{}_1(x) = c$, $(f^a{}_1)^b{}_1(x) = c + 3k$, $k > 0$. Every neighbourhood Ω of x contains points x' where $f^a{}_1(x') > c + 2k$ and therefore neighbourhoods Ω' of x' where the sets of points x'' satisfying $f(x'') > c + k$ has the frequency F_a. Therefore $f^a{}_1(x) > c$, contrary to the supposition. Hence the theorem.

Formula (9) is a special case of the theorem, as it corresponds to $F_a = F_b$. By putting $F_b = F_0$ we obtain:

COROLLARY 1. For every F, the functions $f_1(x)$ and $f_2(x)$ are semi-continuous above and below, respectively.

From a well known theorem[4] therefore follows:

COROLLARY 2. If $\sum$ is an n-dimensional Euclidean space, then $f_1(x)$ and $f_2(x)$ are measurable functions.

This corollary admits generalization to other spaces.

That $(f^a{}_1)^b{}_1(x)$ and $(f^b{}_1)^a{}_1(x)$ may be different functions, can be shown by the following example.

$$f(x) \begin{cases} = 1, & x \text{ a rational number,} \\ = 0, & x \text{ an irrational number.} \end{cases}$$

$$\left.\begin{aligned} (f^0{}_1)^1{}_1(x) &= 1 \\ (f^1{}_1)^0{}_1(x) &= f^1{}_1(x) = 0 \end{aligned}\right\} \text{ for every } x.$$

We call the functions $f(x)$ for which $f(x) = f_1(x)$ (or $f(x) = f_2(x)$), *semi-continuous* mod F *above* (or *below*), generalizing ordinary semicontinuity which corresponds to $F = F_0$. The semicontinuity mod F implies also the corresponding semicontinuity mod every weaker frequency. Addition of a continuous function and multiplication with a positive continuous function leave semicontinuity mod F invariant. Multiplication with -1 interchanges semicontinuity above and below (mod F).

[4] [1] p. 403.

The obvious inequalities:

$$(f(x) + g(x))_1 \leqq f_1(x) + g_1(x),$$
$$(f(x) + g(x))_2 \geqq f_2(x) + g_2(x)$$

can be replaced by the corresponding equalities when $F = F_0$, but not in the general case (example $f(x) = 1$ for $0 \leqq x \leqq 1$, otherwise $f(x) = 0$; $g(x) = f(x+1)$). Semicontinuity above mod F_0 can be tested by the necessary and sufficient condition that $x_i \to x_0$ and $f(x_i) \to a$ imply $f(x_0) \geqq a$; however an arbitrary frequency has no test involving convergence of sequences only.

4. To investigate the functions obtained by applying alternatively the upper and lower limiting operations mod a given F, a more general way of approach is convenient.

Given an arbitrary set A which contains a partially ordered subset A'; let L_1 and L_2 be two mappings of the elements $a, b, \ldots$ of A on elements of A'

$$L_1 : a \to a_1, \quad L_2 : a \to a_2 \tag{11}$$

which satisfy the following conditions, when i, j, k, stand for 1 and 2:

$$a_{ii} = a_i \qquad \text{(idempotent)}, \tag{12}$$

$$a_j \prec b_k \text{ implies } a_{ji} \prec b_{ki} \text{ (monotonic)} \tag{13}$$

$$a_2 \prec a_1. \tag{14}$$

Then $a_{i_1 \ldots i_m 2 k_1 \ldots k_n} \prec a_{i_1 \ldots i_m 1 k_1 \ldots k_n}$. In particular,

$$a_{1212} \prec a_{1112} = a_{12} = a_{1222} \prec a_{1212},$$

and therefore

$$a_{1212} = a_{12}, \; a_{(121)(121)} = a_{12121} = a_{121}. \tag{15}$$

Hence the operations L_{12} mapping $a \to a_{12}$ and L_{121} mapping $a \to a_{121}$ are idempotent and obviously monotonic. The same statements hold for the mappings L_{21} and L_{212} which are defined in a corresponding manner. The six mappings $L_1, L_2, L_{12}, L_{21}, L_{121}, L_{212}$ form a semigroup of idempotents, and the four mappings $L_{12}, L_{121}, L_{21}, L_{212}$ form a subsemigroup in it. We have

$$\begin{aligned} &a_2 \prec a_{212} \prec a_{12} \prec a_{121} \prec a_1, \\ &a_2 \prec a_{212} \prec a_{21} \prec a_{121} \prec a_1. \end{aligned} \tag{16}$$

These formulae do not establish any order relation between a_{12} and a_{21}. By the mappings L_1, L_2 and L_{12}, and with the help of (16), we obtain easily:

$$a_{12} \prec a_{21} \text{ implies } a_{21} = a_{121} \text{ and } a_{212} = a_{12}, \tag{17}$$

$$a_{12} \prec b_{121} \prec a_{121} \text{ implies } a_{121} = b_{121} \text{ and } a_{12} = b_{12}. \tag{18}$$

5. We apply now the methods and results of **4** to the space $\sum$ considered in **3**. The system A consists of the pairs $\{f, x\}$ represented by $f(x)$, where x runs over

C and f over the functions with domain C; the system A' consists correspondingly of the $f_i(x)$; the order relation in A' is the relation $\prec$ introduced by (7), and the mappings L_i are defined by

$$L_i : f(x) \to f_i(x), \quad i = 1,2. \tag{19}$$

From (16) it follows that for every $x \in C$,

$$f_2(x) \prec f_{212}(x) \prec f_{21}(x) \prec f_1(x), \tag{20}$$

$$f_2(x) \prec f_{12}(x) \prec f_{121}(x) \prec f_1(x). \tag{21}$$

It may be remembered that $g(x_0) \prec f(x_0)$ does not necessarily imply $g(x_0) \leqslant f(x_0)$, (nor does $g(x_0) \sim f(x_0)$ imply $g(x_0) = f(x_0)$), but the relation implies $g(x) \leqslant f(x)$ for nearly every x of a suitable neighbourhood Ω of x_0 (similarly for $\sim$). However, by Theorem 3, $g_i(x_0) \prec f_i(x_0)$ implies $g_i(x) \leqslant f_i(x)$ for every $x \in \Omega$. In the case of functions with several suffixes, it is the last one that matters. For $x \in \Omega$,

$$\begin{aligned} &f_{12}(x_0) \prec f_{21}(x_0) && \text{implies } f_{21}(x) = f_{121}(x) \text{ and } f_{212}(x) = f_{12}(x), \\ &f_{12}(x_0) \prec g_{121}(x_0) \prec f_{121}(x_0) && \text{implies } f_{121}(x) = g_{121}(x) \text{ and } f_{12}(x) = g_{12}(x). \end{aligned} \tag{22}$$

If $f(x)$ is continuous at $x = x_0$, then $f_1(x_0) = f_2(x_0)$. Conversely this equation does not necessarily imply the continuity of $f(x)$ at x_0. If for $x \in \Omega$ (open), $f_1(x) = f(x) = f_2(x)$, then (by Corollary 1 of Theorem 4) $f(x)$ is semicontinuous above and below and therefore continuous. Furthermore we prove

THEOREM 5. *Let $f_1(x) \sim f_2(x)$ for $x \in \Omega$ (open), then there exists $K \subseteq \Omega$ such that K contains nearly all the points of Ω and $f(x)$ is continuous when considered as a function with the domain K.*

Proof. $f_2(x) = f_1(x)$, for nearly every $x \in \Omega$; moreover, by Theorem 2, $f_2(x) \leqslant f(x) \leqslant f_1(x)$ for nearly every $x \in C$. Hence the points of Ω which satisfy both these conditions form a set $K \leqslant \Omega$, where $f_2(x) = f(x) = f_1(x)$. Thus $f(x)$ is continuous when considered on K alone.

The theorem admits a converse statement, since the values of $f_1(x)$ and $f_2(x)$ do not depend on the values of $f(x)$ on the complement of K in Ω. Therefore the continuity of $f(x)$ on K implies the equivalence of $f_1(x)$ and $f_2(x)$, $x \in \Omega$. The integrability (C) of a function does not depend on its values on a set which has not the frequency F_C. Hence:

COROLLARY. When $F = F_C$ and for $x \in \Omega$, $f_1(x) \sim f_2(x)$ and $f(x)$ is bounded, then $f(x)$ is integrable (C) and $\int f(x)\,dx = \int f_1(x)\,dx$.

It should be noticed that we have here a sufficient condition for integrability (C) which depends on a property "im Kleinen" only. For a Euclidean space $\sum$ and Lebesgue integration, the class of functions satisfying the conditions of the corollary does not include all the bounded functions which are measurable (L), but is larger than the class of the functions which are bounded and integrable (R).

THEOREM 6. *Let $g(x)$ be continuous on the open set Ω, let A be the subset of Ω where $g(x) \geqslant f_{121}(x)$ and B the subset where $g(x) \leqslant f_{12}(x)$; then $A \cup B$ has the frequency F at every point of Ω.*

Proof. Suppose $A \cup B$ has not the frequency F at $x_0 \in \Omega$; then $f_{12}(x_0) < g(x_0) < f_{121}(x_0)$. As $g(x)$ is continuous, $g(x) = g_{12}(x) = g_{121}(x)$. Therefore it follows from (22) that $g(x)$ is equal and equivalent to $f_{12}(x)$ and to $f_{121}(x)$ at x_0. Hence $x_0 \in A \cap B \subseteq A \cup B$. At the points of the complement C of $A \cup B$ in Ω, therefore, $A \cup B$ has the frequency F. This leads to a contradiction, since when $A \cup B$ has not the frequency F at x_0, this point is a limiting point of C and therefore $A \cup B$ has the frequency F at x_0.

6. Consider now *pairs* of functions $g(x)$, $h(x)$ for which, for $x \in C$,

$$g(x) = g_1(x) = h_1(x) \text{ and } h(x) = g_2(x) = h_2(x); \tag{23}$$

then

$$\begin{aligned} g(x) &= g_{121}(x) = h_{121}(x) = g_{21}(x) = h_{21}(x); \\ h(x) &= g_{12}(x) = h_{12}(x) = g_{212}(x) = h_{212}(x). \end{aligned}$$

On the other hand, for every function $f(x)$, the pairs $f_{121}(x)$, $f_{12}(x)$ and $f_{21}(x)$, $f_{212}(x)$ satisfy (23). Now suppose

$$h(x_0) < f(x_0) < g(x_0); \tag{24}$$

then $g(x_0) = h_1(x_0) \leqslant f_1(x_0) \leqslant g_1(x_0) = g(x_0)$; hence $f_1(x_0) = g(x_0)$. Similarly $f_2(x_0) = h(x_0)$. On the other hand, $f_1(x_0) = g(x_0)$ implies (see Theorem 2) $f(x_0) < g(x_0)$, and if $f_1(x) = g(x)$ for every $x \in \Omega$ (open set), then $f(x) \leqslant g(x)$ for nearly every $x \in \Omega$. Thus:

THEOREM 7. *Let the pair of functions $g(x)$, $h(x)$ satisfy* (23); *then the necessary and sufficient condition for $f_1(x_0) = g(x_0)$, $f_2(x_0) = h(x_0)$ is* (24).

When Ω is an open set, the necessary and sufficient condition for $f_1(x) = g(x)$, $f_2(x) = h(x)$, $x \in \Omega$ is $h(x) < f(x) < g(x)$.

We consider now two frequencies $F_a \subseteq F_b$ which $\sum$ has at every point. To avoid clumsy formulas, we use the indices 1,2 mod F_b and correspondingly the indices α, β mod F_a. If $g(x)$ and $h(x)$ satisfy (23), then it follows from (10) that

$$h(x) = g_2(x) \leqslant g_\beta(x) \leqslant g_\alpha(x) \leqslant g_1(x) = g(x),$$

but from Theorems 4 and 7, $g_\alpha(x) = g_{\alpha 1}(x) = g(x)$,

$$g_\beta(x) = g_{\beta 2}(x) = h(x).$$

Now suppose that $r(x) = r_\alpha(x) = s_\alpha(x)$, $s(x) = s_\beta(x) = r_\beta(x)$, then $r_1(x) = r(x)$, $s_2(x) = s(x)$ and therefore $s(x) = r_\beta(x) \geqslant r_2(x) = r_{\beta 2}(x) = s_2(x) = s(x)$. Hence $r_2(x) = s(x)$, and similarly $s_1(x) = r(x)$. Therefore if (23) holds for $F = F_b$, it holds also for F_a, and, conversely, if it holds for any F, it holds for F_0 and for every other F' (as it is necessarily $\subseteq F_0$). Hence

THEOREM 8. *If F and F' are frequencies which C has at every point, and the equations* (23) *are satisfied* mod F, *they are also satisfied* mod F'.

In the supposition of Theorem 8, C may be replaced by any subspace. Moreover it follows from this theorem, that when A and B are defined as in Theorem 6, $A \cup B$ has every frequency that $\Omega \cap C$ has at each of its points.

COROLLARY.[5] For every $\epsilon > 0$ and arbitrary x_0, the set of points x for which $g(x) \geqslant g(x_0) - \epsilon$ has all the frequencies at x_0 which C has at that point.

As the relation (23) between $g(x)$ and $h(x)$ does not depend on the selection of F, we may put $F = F_0$; therefore $g(x_0) = h(x_0)$ is the necessary and sufficient condition for $g(x)$ $(= h(x))$ to be continuous at x_0. The difference

$$\delta(x) = g(x) - h(x) \tag{25}$$

can therefore be used as a measure of the discontinuity of $g(x)$ and $h(x)$. $\delta(x)$ is non-negative and semicontinuous above mod F_0, but not every function with these properties is a $\delta(x)$. Select an arbitrary F which C has at every point, then $\delta(x) \geqslant \delta_1(x)$ (for (10) holds nearly everywhere). Suppose $\epsilon > 0$; then there exists a neighbourhood Ω of x_0, such that for $x \in \Omega \cap C$, $g(x) \leqslant g(x_0) + \epsilon$; moreover, there exists a subset $S \subseteq \Omega \cap C$ which has the frequency F at x_0, such that for $x' \in S$, $h(x') \geqslant h_1(x_0) - \epsilon = g(x_0) - \epsilon$. Hence $\delta(x') \leqslant 2\epsilon$ for $x' \in S$. As $\delta(x) \geqslant 0$ and S has the frequency F, it follows that $\delta_2(x_0) = 0$ for every x_0. Moreover $\delta_{12}(x) \leqslant \delta_2(x)$. Hence $0 = \delta_{12}(x) = \delta_{121}(x) = \delta_{21}(x) = \delta_{212}(x) = \delta_2(x)$. Therefore $\delta(x)$ is continuous only at those points where it vanishes, i.e., where $g(x) = h(x)$ is continuous.

The pairs $g(x)$, $h(x)$ which satisfy (23) are the same for every F; but for a given function $f(x)$, the pairs $f_{121}(x)$, $f_{12}(x)$ are in general different for different frequencies F.

7. A discontinuous function $f(x)$ can be characterized by the two semicontinuous functions $f_1(x)$ and $f_2(x)$, and furthermore by the two "pairs" $f_{121}(x)$, $f_{12}(x)$ and $f_{21}(x)$, $f_{212}(x)$. This characterization depends on the choice of the frequency F. It is therefore interesting to know all the frequencies which the domain of definition of $f(x)$ has at each of its points. We have mentioned frequencies which are derived from the powers of the subsets of $\sum$ and frequencies derived from measure functions. To know all the frequencies of the first kind would imply the solution of the "problem of the continuum"; the frequencies of the second kind include those generated by measures of lower dimension (e.g., Gillespie measure).[6] Moreover, if F is a frequency which $T \subseteq \sum$ has at every point of $\sum$, then $\sum$ has also the frequency $F(T)$ at every point. However there might exist frequencies of a different kind.

[5] If one tries to describe the domain of the values between $g(x)$ and $h(x)$, spread over C in a pictorial way, this corollary states that the domain is "soft" inside, whereas Theorem 6 means that it is "hard" outside.

[6] See [3].

The functions $f_1(x)$ and $f_2(x)$ satisfy $f_2(x) \leqslant f_{12}(x)$. To show that there is no other relation between those two semicontinuous functions, we select an arbitrary function which is semicontinuous above, say $r(x)$, and a function $s(x)$, semicontinuous below, such that $s(x) \leqslant r_2(x)$ for $x \in \sum$. Then we split $\sum$ into $\sum = A \cup B$, $A \cap B = 0$, such that A as well as B has the frequency F at every point of $\sum$. We define

$$f(x) \begin{cases} = r(x) \text{ for } x \in A, \\ = s(x) \text{ for } x \in B; \end{cases}$$

then $f_1(x) = r(x)$, $f_2(x) = s(x)$, $x \in \sum$.

That the splitting of $\sum$ into A and B is always possible for $F = F_0$ follows from the separability of $\sum$. To split the n-dimensional Euclidean space into two sets A and B which have a positive Lebesgue measure at each point of the space, one can construct a suitable sequence of disjoint perfect sets of positive measure $A_1, B_1, A_2, B_2, \ldots$ such that $A = \cup_n A_n$ and $B = \cup_n B_n$ have the required property.

Thus there is no relation other than $f_2(x) \leqslant f_{12}(x)$, between the upper and the lower limiting function, which holds for every frequency; and therefore a closer investigation of the nature of the discontinuities must be split into the "discontinuity above" (characterized by $f_1(x)$ and the pair $f_{121}(x), f_{12}(x)$) and the "discontinuity below" ($f_2(x)$ and $f_{21}(x), f_{212}(x)$).

It has been suggested[7] that we might modify the notion of upper (lower) limiting function by considering the functions

$$D_1(f,x_0) = \overline{\lim_{x \to x_0}} f(x); \; D_2(f,x_0) = \underline{\lim_{x \to x_0}} f(x).$$

This would be an Analysis modulo the distributive property D: "To contain an infinite number of points". These limiting functions, however, do not lead to an idempotent operation, not even after infinite repetition, as is seen from the following example:

Represent the numbers $0 \leqslant x < 1$ by decimal fractions; then $x \in S_{m,n}$ $(n \leqslant m)$ if and only if either $x = 0$, or x admits a finite decimal expansion which starts with *exactly* m zeros and has altogether $m - n$ non-zeros. Put

$$\cup_m \{S_{m,0} \ldots S_{m,k}\} = S^k, \; \cup_k S^k = S,$$

and define $f(x) = 1$ for $x \in S$, $= 0$ otherwise; then $D_1(f, x) = 1$ if and only if $x \in S - S^0$, and when we indicate the iteration of the D_1-operation by an upper index, $D_1^{k+1}(f, x) = 1$ if and only if $x \in S - S^k$. All these operators are therefore non-idempotent, and the functions form a monotonic decreasing sequence. The lower limit is $D_\omega(f, x)$ which $= 1$ for $x = 0$ only; thus the operator D_ω is not idempotent either. The example can be modified in such a way that even some operators with higher transfinite indices are non-idempotent. For this purpose, one may admit several batches of non-zeros separated by sequences of consecutive zeros of prescribed length.

[7] See [4] p. 1003, footnote 452a.

8. Given a frequency F which $C \subseteq \sum$ has at every point of C, and a sequence $f(x,1), f(x,2), f(x,3), \ldots$ which converges to a function $f(x)$ defined on a subset of C in such a way that for every $\epsilon > 0$ there exists an $N(\epsilon)$, such that for $n > N(\epsilon)$,

$$|f(x,n) - f(x)| < \epsilon \quad \text{for nearly every } x \in C, \tag{26}$$

then the sequence is said to converge *nearly uniformly* to $f(x)$ on C. Let $\epsilon_i \to 0$; if (26) holds, then the corresponding inequality holds also for every $\epsilon_i > \epsilon$. If $C_{n,i}$ is the set of points x for which

$$|f(x,n) - f(x)| \geqslant \epsilon_i,\ n > N(\epsilon_i)$$

then $\bigcup_{n,i} C_{n,i}$ has not the frequency F and therefore $B = C - \bigcup_{n,i} C_{n,i}$ has the frequency F at every point of C. Therefore the given sequence is uniformly convergent on B to $f(x)$; this function is defined at every point of B, and its limiting functions $f_1(x), f_2(x)$ are defined for every $x \in C$. We prove now:

THEOREM 9. (26) *implies that* $f_i(x,n)$ *converge to* $f_i(x)$ *uniformly on* C *for* $i = 1,2$.

Proof. Let U_n be a suitable neighbourhood of $x_0 \in C$. Then $f(x,n) < f_1(x_0,n) + \epsilon$ for nearly every $x \in U_n \cap B$, and therefore

$$f(x) < f_1(x_0,n) + 2\epsilon,\ n > N(\epsilon).$$

Moreover, U_n has a subset K_n with the frequency F, such that $f(x',n) > f_1(x_0,n) - \epsilon$ and therefore

$$f(x') > f_1(x_0,n) - 2\epsilon,\ n > N(\epsilon),\ x' \in K_n.$$

Now $f_1(x_0)$ exists for $x_0 \in C$, and it follows from the two inequalities that

$$f_1(x_0,n) - 2\epsilon \leqslant f_1(x_0) \leqslant f_1(x_0,n) + 2\epsilon,\ n > N(\epsilon).$$

Therefore $f_1(x_0,n) \to f_1(x)$ and similarly $f_2(x_0,n) \to f_2(x)$. Moreover,

$$|f_1(x_0) - f_1(x_0,n)| < 2\epsilon,\ n > N(\epsilon) \text{ independent of } x_0.$$

Thus the convergence is uniform on C.

REFERENCES

[1] C. Caratheodory, *Vorlesungen über reelle Funktionen* (B. G. Teubner, 1919).
[2] F. W. Levi, *On the Fundamentals of Analysis* (Calcutta, 1939).
[3] A. P. Morse and J. F. Randolph, "Gillespie Measure," *Duke Math. J.*, vol. 6, pp. 408-419.
[4] Zoretti-Rosenthal, "Die Punktmengen," *Encyklopädie der mathematischen Wissenschaften*, II C9a.

Tata Institute of Fundamental Research,
Bombay

THE FACTORIZATION OF LOCALLY FINITE GRAPHS

W. T. TUTTE

1. Introduction. A *graph* G consists of a set $V(G)$ of objects called *nodes* and a set $M(G)$ of objects called *links*, $V(G)$ and $M(G)$ having no members in common. With each link A there is associated just two nodes said to be the *ends* of A, or to be *incident* with A, or to be *joined* by A. The sets $V(G)$ and $M(G)$ may be finite or infinite. There may be nodes with which no link is incident. Such nodes are said to be *isolated*.

If $V(G)$ and $M(G)$ are finite, the graph G is said to be *finite*.

The *order* of a graph G is the cardinal of $V(G)$. The *degree* of a node a of G is the cardinal of the set of links of G with which a is incident. The graph G is said to be *locally finite* if the degree of each node of G is finite. If all the nodes of G have the same finite degree σ we may say that G is a *regular* graph of the σth *degree*.

Let x and y be any two nodes of a graph G. We say that they are *connected* in G if there exists a finite sequence

$$P = (b_1, B_1, b_2, B_2, \ldots, b_r, B_r, b_{r+1}) \tag{1}$$

satisfying the following conditions.

(i) $b_1 = x$; $b_{r+1} = y$.

(ii) The members of P are alternately nodes and links of G.

(iii) Consecutive members of P are incident.

If x and y are any nodes of a graph G, we define the relation $x \backsim y$ to mean that either $x = y$ or else x and y are connected in G. It is easily verified that the relation is an equivalence relation. It therefore partitions G into a set $\{G_\alpha\}$ of graphs such that G_α is connected, each node or link of G belongs to some G_α, and no two of the G_α have any node or link in common. We shall call the G_α the *components* of G.

A *subgraph* of a graph G is a graph G' such that $V(G') = V(G)$ and $M(G') \subseteq M(G)$, a link of G' having the same ends in G' as in G. A *factor* of G is a regular subgraph of G of the first degree. If G has no factor it is *prime*. Clearly all finite graphs of odd order are prime.

Let S be any finite subset of $V(G)$. Then we denote the number of members of S by $f(S)$. We denote by G_S the graph obtained from G by suppressing the members of S and all links of G having one or both ends in S. Let $h(S)$ be the cardinal of the set of components of G_S, and let $h_u(S)$ be the cardinal of the set of those components of G_S which are finite and of odd order. Clearly, if G is locally finite and is connected, then $h(S)$ and $h_u(S)$ are finite.

The object of this paper is to prove the following Theorem.

Received October 4, 1948.

THEOREM A. *A locally finite graph G is prime if and only if there is a finite subset S of $V(G)$ such that $h_u(S) > f(S)$.*

A proof of this Theorem for the case in which G is finite has already been given. (W. T. Tutte, "The Factorization of Linear Graphs," *J. London Math. Soc.*, vol. 22 (1947), 107-111). We refer to this paper below as Paper I. In the present paper we assume the truth of the Theorem for finite graphs and show how to extend it to the case in which G is infinite (but locally finite).

2. Preliminary results. We shall say that a graph G is *constricted* if there exists a finite subset S of $V(G)$ such that $h_u(S) > f(S)$.

THEOREM I. *A constricted graph has no factor.*

Let G be any graph such that $V(G)$ has a finite subset S such that $h_u(S) > f(S)$.

Suppose there exists a factor F of G. Then if C is any finite component of odd order of G_S it is clear that F must contain a link having one end in C and the other in S. Hence the cardinal of the set of links of F having ends in S is greater than the number of members of S. Hence some node of S must be incident with more than one link of F, which is absurd.

THEOREM II. *The truth of Theorem A for connected locally finite graphs implies its truth for all locally finite graphs.*

Let G be any locally finite graph, and let $\{G_\alpha\}$ be the set of its components. Let us assume that Theorem A has already been proved for connected locally finite graphs.

If G is prime, some component G_α of G must be prime. For if each G_α had a factor F_α we could clearly obtain a factor of G by combining the factors F_α.

But if G_α is prime there is a finite subset S of $V(G_\alpha)$ such that $f(S) < h_u(S)$ in G. For G_α is connected and so we can apply Theorem A to it. But the components of G_S are simply the components of $(G_\alpha)_S$ together with the components other than G_α of G. Hence the inequality $f(S) < h_u(S)$ is true also in G.

Hence if G is prime, it is constricted. But by Theorem I, if G is constricted, it is prime. Thus Theorem A is true for G.

We conclude from Theorems I and II that, in order to prove Theorem A, it now suffices to prove that any connected locally finite infinite graph which is not constricted has a factor.

3. Distance and n-factors. Let G be any connected locally finite graph, and let a be any node of G.

Suppose that b is any other node of G. Then because G is connected there exists a finite sequence P of the form (1) such that $b_1 = a$ and $b_{r+1} = b$. The least value of r for which such a sequence P exists will be called the *distance* $d(a, b)$ from a to b. We write $d(a, a) = 0$.

We define $V_n(G; a)$ to be the set of all nodes b of G such that $d(a, b) \leq n$. We define $M_n(G; a)$ to be the set of all links A of G such that both ends of A are in $V_n(G; a)$. It is clear from these definitions that if a link A of G has one

end in $V_n(G; a)$, then $A \in M_{n+1}(G; a)$. It is also evident, from the fact that G is locally finite, that $V_n(G; a)$ and $M_n(G; a)$ are finite for each non-negative integer n.

We define a *false factor* of G as a subgraph X of G which satisfies the following conditions.

(i) $M(X)$ is finite.

(ii) There is no node of G whose degree in X exceeds 1. If X is a false factor of G, we denote by $W(X)$ the set of all nodes of X which are incident with links of X.

If a is a node and X a false factor of G, we say that X is an *n-factor of G with respect to a* if

$$V_{n+1}(G; a) \supseteq W(X) \supseteq V_n(G; a), \tag{2}$$

n being any non-negative integer.

THEOREM III. *Let G be any connected locally finite infinite graph which is not constricted, and let a be any node of G. Let n be any non-negative integer. Then there exists an n-factor of G with respect to a.*

Let Q denote the set $V_{n+1}(G; a) - V_n(G; a)$.

We define a graph H as follows. The nodes of H are the members of $V_{n+1}(G; a)$. We take each member of $M_{n+1}(G; a)$ as a link of H, assigning it the same ends in H as in G. In addition we join each pair of members of Q by a new link.

We next define a graph K. If the order of H is even, $K = H$. If the order of H is odd we construct K from H by adjoining to H a new node q and then joining q to each member of Q by a new link. By this construction the order $m(K)$ of K is always even. We write $Q' = Q$ or $Q \cup \{q\}$ according as the order of H is even or odd.

Suppose that K is constricted. Then there is a subset S, (possibly the null subset), of $V(K)$ such that

$$h_u(S) > f(S) \text{ in } K. \tag{3}$$

But it is clear that

$$m(K) \equiv h_u(S) + f(S) \pmod 2. \tag{4}$$

Consequently, since $m(K)$ is even, we must have

$$h_u(S) \geq f(S) + 2. \tag{5}$$

At most one of the components of K_S contains a node of Q'. For any two nodes of Q' will be joined by a link of K. Two such nodes cannot therefore be in different components of K_S. Write $T = S$ or $S - \{q\}$ according as the order of H is even or odd. (If $q \bar{\in} S$, $S - \{q\} = S$). Then it is clear that any component of K_S which contains no node of Q' is also a component of G_T. Hence, using (5), we deduce that the inequality $h_u(T) > f(T)$ holds in G. This is contrary to our hypothesis that G is not constricted. We conclude that K is not constricted.

Since Theorem A holds for finite graphs it follows that K has a factor F_1. Let X be the subgraph of G defined by

(6) $$M(X) = M(F_1) \cap M_{n+1}(G; a).$$

Each node of G is incident with at most one member of $M(F_1)$. Hence X is a false factor of G. Now each link of K incident with any node $c \epsilon V_n(G; a)$ is a link of $M_{n+1}(G; a)$ incident with c in G. Hence c is incident in G with a member of $M(X)$. Thus

(7) $$W(X) \supseteq V_n(G; a).$$

Also each node of G incident with a member of $M_{n+1}(G; a)$ is a member of $V_{n+1}(G; a)$. Hence

(8) $$V_{n+1}(G; a) \supseteq W(X).$$

From (7) and (8) it follows that X is an n-factor of G with respect to a.

4. Proof of Theorem A. In this Section, G is any connected, locally finite, infinite graph, which is not constricted, and a is any node of G.

Let m and n be integers satisfying $m \geq n \geq 0$, and let X_m be any m-factor of G with respect to a. We denote by $C(X_m; n)$ the subgraph of G obtained from X_m by suppressing all links of X_m not in $M_{n+1}(G; a)$. It is clear that if $l \geq n$, then

(9) $$C(C(X_l; m); n) = C(X_l; n).$$

It is also evident that $C(X_m; n)$ is an n-factor of G with respect to a.

Suppose that the m-factor X_m and the n-factor X_n of G with respect to a are related by the equation

$$X_n = C(X_m; n).$$

Then we say that X_m is an *extension* of X_n to m. It may happen for a given n-factor X_n of G with respect to a, that there exists an integer $m > n$ such that X_n has no extension to m. In that case it follows by (9) that X_n has no extension to any integer $l > m$. There is thus a maximum integer $r(X_n) \geq n$ such that X_n has an extension to $r(X_n)$. We call $r(X_n)$ the *range* of X_n. The only other possibility is that the given n-factor X_n may have extensions to all integers $m \geq n$. We then say that X_n has *infinite range.*

THEOREM IV. *There exists a 0-factor of G with respect to a which has infinite range.*

If X_0 is any 0-factor of G with respect to a, it follows from (2) that $M(X_0) \subseteq M_1(G; a)$. As $M_1(G; a)$ is finite, it follows that the set of 0-factors of G with respect to a is finite.

Suppose that no one of them has infinite range. Then there exists an integer n greater than the range of any 0-factor of G with respect to a. By Theorem III there exists an n-factor X_n of G with respect to a. Then $C(X_n; 0)$ is a 0-factor of G with respect to a whose range is not less than n. This contradiction establishes the Theorem.

THEOREM V. *Let n be any non-negative integer, and let X_n be an n-factor of G with respect to a having infinite range. Then there exists an $(n+1)$-factor X_{n+1} of G with respect to a which has infinite range and which satisfies $X_n = C(X_{n+1}; n)$.*

Let Z be the set of $(n+1)$-factors of G with respect to a which are extensions to $(n+1)$ of X_n. If X is any member of Z it follows from (2) that $M(X) \subseteq M_{n+1}(G; a)$. As $M_{n+1}(G; a)$ is finite it follows that Z is finite.

Suppose that no member of Z has infinite range. Then there is an integer $m > n+1$ which exceeds the range of each member of Z. Since X_n has infinite range there exists an m-factor X_m of G with respect to a such that $C(X_m; n) = X_n$. Then $C(X_m; n+1)$ is an $(n+1)$-factor of G with respect to a. By (9) it is a member of Z. Hence it can have no extension to m. This is absurd, since X_m is such an extension. The Theorem follows.

THEOREM VI. *There exists an infinite sequence* $(Y_0, Y_1, Y_2, Y_3, \ldots)$ *having the following properties.*

(i) *For each non-negative integer* n, Y_n *is an* n*-factor of* G *with respect to* a *having infinite range.*

(ii) *For each non-negative integer* n, $C(Y_{n+1}; n) = Y_n$.

The terms Y_r are defined successively as follows. Y_0 is defined to be a 0-factor of G with respect to a having infinite range. Such a Y_0 exists, by Theorem IV. For $r \geq 0$, Y_{r+1} is defined to be an extension to $(r+1)$ of Y_r having infinite range. If Y_r is fixed, such a Y_{r+1} exists, by Theorem V. The sequence of Y_r defined in this way satisfies (i) and (ii).

Let us consider some particular infinite sequence $(Y_0, Y_1, Y_2, \ldots)$ satisfying conditions (i) and (ii) of Theorem VI. Using (9) we can show, by an obvious induction, that if m and n are integers satisfying $m \geq n \geq 0$, then $C(Y_m); n) = Y_n$.

Let F be the subgraph of G defined by

$$M(F) = \bigcup_{r=0}^{\infty} M(Y_r). \tag{10}$$

Let b be any node of G. Write $d(a, b) = n$. Then $b \epsilon V_n(G; a) \subseteq W(Y_n)$, by (2). Hence b is incident in G with some link of F. Suppose that b is incident with two distinct links A_1 and A_2 of F. Then by (10) there exist integers s and t, ≥ 0, such that $A_1 \epsilon M(Y_s)$ and $A_2 \epsilon M(Y_t)$. Then $A_1 \epsilon M(Y_u)$ and $A_2 \epsilon M(Y_u)$, where u is any integer greater than s and t. For $C(Y_u; s) = Y_s$ and $C(Y_u; t) = Y_t$. But then the degree of b in the false factor Y_u of G exceeds 1. This contradicts the definition of a false factor of G.

From these considerations we conclude that each node of G has degree 1 in F. That is, F is a factor of G.

It now follows, from Theorem I, that Theorem A holds for all connected, locally finite, infinite graphs. Since, by Paper I, it holds for all finite graphs, we see that it holds for all connected locally finite graphs. Hence by Theorem II it holds for all locally finite graphs.

5. Regular graphs. If G is a connected locally finite graph we define an *isthmoid* of G as a finite subset S of $V(G)$ such that $h(S) > 1$. We then say that $f(S)$ is the *rank* of the isthmoid.

We find that Theorem V of Paper I, and its Corollary, can be generalized as follows.

THEOREM VII. *Let G be any connected locally finite graph which is regular and of degree $\sigma > 0$, and which is either infinite or else of even order. Suppose further that G has no isthmoid of rank $< \sigma - 1$. Then G has a factor.*

COROLLARY. *Let A be any link of G. Then G has a factor which contains A.*

Here we shall only consider the case in which G is infinite, the finite case having been dealt with in Paper I. It will be found that the argument of Paper I remains valid in the infinite case as far as the Theorem is concerned, if we replace the appeal to Theorem IV of Paper I by an appeal to Theorem A. The proof of the Corollary in Paper I is not valid for the infinite case. We may replace it by the following argument (which is not valid for the finite case).

Let x and y be the ends of A. Suppose that the Corollary fails for some graph G. Then $G_{[x,\ y]}$ is prime. Hence, by Theorem A, there exists a finite subset S of $V(G) - \{x, y\}$ such that $h_u(S) > f(S)$ in $G_{[x,\ y]}$.

Let S' be the set formed by adjoining x and y to S. Hereafter functions of S will refer to $G_{[x,\ y]}$ and functions of S' to G. Clearly

$$f(S') = f(S) + 2 \tag{11}$$

and

$$h_u(S') = h_u(S). \tag{12}$$

Now if C is a finite component of $G_{S'}$ of odd order, the number of links of G having one end in C and the other in S' is at least σ. (Paper I, proof of Theorem V). Apart from any such components $G_{S'}$ has at least one infinite component C_∞. For G is infinite and connected, and each node of S' is of finite degree. The number of links of G having one end in C_∞ and the other in S' is at least $\sigma - 1$, since G has no isthmoid of rank less than $\sigma - 1$.

Let k be the number of links of G having just one end in S'. Using the above considerations and the fact that A has both ends in S' we find

$$(\sigma - 1) + \sigma.h_u(S') \leqslant k \leqslant \sigma f(S') - 2.$$

Hence, since $\sigma > 0$,

$$f(S') \geq h_u(S') + 1 + 1/\sigma, \tag{13}$$

whence

$$f(S') \geq h_u(S') + 2, \tag{14}$$

since $f(S')$ and $h_u(S')$ are integers.

It follows from (11), (12) and (14) that $f(S) \geq h_u(S)$. This contradicts the definition of S. The Corollary follows.

University of Toronto

UN THÉORÈME DE TRANSFERT D'UN ANNEAU ABSTRAIT À L'ANNEAU DES POLYNOMES

LÉONCE LESIEUR

CERTAINS théorèmes démontrent pour l'anneau des polynomes $A[x_1, \ldots, x_n]$ une propriété supposée valable dans l'anneau A, par exemple le théorème de la base finie de Hilbert pour tout idéal d'un anneau commutatif qui possède un élément unité. (Voir P. Dubreil [2] p. 210, ou B. L. Van der Waerden [7] §80, ou le mémoire original de Hilbert [3], p. 473.) E. Lasker [5] a montré que tout idéal de l'anneau $K[x_1, x_2, \ldots, x_n]$ des polynomes à n variables à coefficients dans un corps K commutatif est l'intersection d'un nombre fini d'idéaux primaires. Sa démonstration repose sur la théorie de l'élimination. E. Noether ([7] §83) a établi plus généralement le décomposition en un nombre fini d'idéaux primaires pour tout idéal d'un anneau satisfaisant le théorème de la base finie (anneau noetherien).

Le présent travail apporte d'abord, aux paragraphes 1 et 2, un nouveau théorème de transfert d'un anneau A à l'anneau $A[x_1, \ldots, x_n]$ des polynomes à n variables. Je suppose la propriété suivante valables dans A (commutatif et possédant un élément unité).

PROPRIÉTÉ 1. Tout idéal $\mathfrak{i}$, différent de l'idéal unité A, admet au moins un diviseur premier $\mathfrak{p}$, différent de A, tel que pour tout élément p de $\mathfrak{p}$ il existe au moins un entier ρ, et un élément l associé, vérifiant:

$$p^\rho . l \equiv 0(\mathfrak{i}), \text{ avec } l \not\equiv 0 \ (\mathfrak{p}).$$

Cette propriété se conserve alors pour l'anneau $A[x_1, \ldots, x_n]$. (théorème 1).

Je donne ensuite, au paragraphe 3, une application au théorème des zéros de Hilbert (Nullstellensatz) et au théorème exprimant qu'un idéal premier dans $K[x_1, \ldots, x_n]$ est l'idéal associé à sa variété (théorème 6); au paragraphe 4, je présente au moyen d'une démonstration élémentaire la décomposition d'un idéal dans $K[x_1, \ldots, x_n]$ en un nombre fini d'idéaux primaires.

1. Remarques préliminaires. A est un anneau commutatif avec élément unité e. Soit $\mathfrak{J}$ un idéal dans l'anneau $A[x]$ des polynomes à une variable. L'ensemble des polynomes de l'idéal $\mathfrak{J}$ qui se réduisent à des éléments de l'anneau A des coefficients constitue un idéal dans A:

$$\mathfrak{i} = \mathfrak{J} \cap A.$$

C'est l'intersection de $\mathfrak{J}$ avec A, que nous appellerons aussi la *projection*[1] de l'idéal $\mathfrak{J}$ dans A. Lorsque l'idéal $\mathfrak{J}$ n'est pas l'idéal unité $\mathfrak{R} = A[x]$, sa projection $\mathfrak{i}$ n'est pas l'idéal unité A; sinon, l'idéal $\mathfrak{i}$ admettrait e comme

Reçu le 24 Janvier, 1949.

[1]On pourrait encore dire, d'après Krull [4], "restriction".

élément, et de même $\mathfrak{J}$ qui serait confondu avec $\mathfrak{R}$. Quand l'idéal $\mathfrak{J}$ est un idéal premier $\mathfrak{P}$ dans $A[x]$, sa projection i est un idéal premier $\mathfrak{p}$ dans A.

Soit $\mathfrak{p}$ un idéal premier différent de A. L'anneau $A/\mathfrak{p}$ est un domaine d'intégrité dont le corps des quotients k est constitué par les éléments

$$\frac{\overline{u}}{\overline{w}}$$

où $\overline{u}$ et $\overline{w}$ sont les classes de u et w modulo $\mathfrak{p}$, avec $w \not\equiv 0(\mathfrak{p})$.

A chaque polynome

$$F(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 \epsilon A[x]$$

on peut faire correspondre le polynome

$$\overline{F(x)} = \overline{a}_n x^n + \overline{a}_{n-1}x^{n-1} + \ldots + \overline{a}_0 \epsilon k[x].$$

Cette correspondance est un homomorphisme λ de $A[x]$ dans $k[x]$.

Quand $F(x)$ décrit l'idéal $\mathfrak{J}$ dans $A[x]$, les polynomes de la forme

$$\frac{\lambda F(x)}{\overline{w}}, \text{ où } w \not\equiv 0(\mathfrak{p}) \text{ dans } A$$

constituent un idéal $\overline{\mathfrak{J}}$ dans $k[x]$. Posons

$$\overline{\mathfrak{J}} = \overline{\lambda}(\mathfrak{J}).$$

La correspondance $\overline{\lambda}$ ramène l'étude de certaines propriétés des idéaux dans $A[x]$ à des propriétés simples des idéaux bien connus dans $k[x]$. Elle respecte évidemment l'inclusion des idéaux dans $A[x]$, c'est à dire:

$$\mathfrak{J} \subseteq \mathfrak{F} \text{ entraine } \overline{\mathfrak{J}} \subseteq \overline{\mathfrak{F}}.$$

De plus, elle a les propriétés du lemme suivant, qui nous sera utile:

LEMME 1. *Lorsque $\mathfrak{P}$ est un idéal premier différent de $\mathfrak{R} = A[x]$, tel que $\mathfrak{P} \cap A = \mathfrak{p}$, l'idéal $\mathfrak{P} = \overline{\lambda}(\mathfrak{P})$ est un idéal premier dans $k[x]$, différent de l'anneau $k[x]$.*

Inversement, $\overline{\mathfrak{P}}$ étant un idéal premier différent de l'anneau $k[x]$, il n'existe qu'un idéal premier $\mathfrak{P}$ dans $A[x]$, vérifiant

$$\mathfrak{P} \cap A = \mathfrak{p} \text{ et } \overline{\lambda}(\mathfrak{P}) = \overline{\mathfrak{P}}.$$

Cet idéal, différent de $\mathfrak{R}$, est constitué par les polynomes $F(x)$ tels que

$$\lambda F(x) \epsilon \overline{\mathfrak{P}}.$$

Démonstration: soit $\mathfrak{P}$ un idéal premier, différent de $\mathfrak{R}$, qui se projette dans A suivant $\mathfrak{p}$. Montrons que $\overline{\mathfrak{P}}$ est premier. Supposons

$$K_1(x).\, K_2(x) \equiv 0 \quad (\overline{\mathfrak{P}}).$$

Les polynomes $K_1(x)$ et $K_2(x)$ peuvent s'écrire

$$K_1(x) = \frac{\lambda A_1(x)}{\overline{w}_1}; \; K_2(x) = \frac{\lambda A_2(x)}{\overline{w}_2}$$

avec $\qquad A_1(x) \quad \text{et} \quad A_2(x) \epsilon A[x].$

D'où, d'après la définition de $\overline{\mathfrak{P}}$

$$\frac{\lambda A_1(x)}{\overline{w}_1} \cdot \frac{\lambda A_2(x)}{\overline{w}_2} = \frac{\lambda P(x)}{\overline{w}} \quad \text{avec } P(x) \epsilon \mathfrak{P}.$$

On en déduit dans $A[x]$:

$$w A_1 A_2 = w_1 w_2 P + \Pi(x),$$

$\Pi(x)$ étant un polynome à coefficients dans $\mathfrak{p}$. Le deuxième membre appartient à $\mathfrak{P}$ puisque $\mathfrak{p} \subset \mathfrak{P}$. La condition $w \not\equiv 0(\mathfrak{p})$ entraine $w \not\equiv 0(\mathfrak{P})$ puisque $\mathfrak{P} \cap A = \mathfrak{p}$. Comme $\mathfrak{P}$ est premier, l'un des polynomes $A_1(x)$ ou $A_2(x)$ appartient à $\mathfrak{P}$. Donc l'un des polynomes $K_1(x)$ ou $K_2(x)$ appartient à $\overline{\mathfrak{P}}$, qui est bien un idéal premier. Démontrons que $\overline{\mathfrak{P}}$ est différent de $k[x]$. Si l'idéal $\overline{\mathfrak{P}}$ remplissait $k[x]$, il contiendrait une constante non nulle de k, soit

$$\frac{\overline{u}'}{\overline{w}'} = \frac{\lambda F_0(x)}{\overline{w}}$$

avec

$$F_0(x) = a_n x^n + \ldots + a_0 \epsilon \mathfrak{P}$$

et

$$\bar{a}_n = 0, \ldots, \bar{a}_1 = 0; \bar{a}_0 \neq 0$$

ou

$$a_n \equiv \ldots \equiv a_1 \equiv 0(\mathfrak{p}); a_0 \not\equiv 0(\mathfrak{p}).$$

Mais on tire de l'égalité

$$a_0 = F_0(x) - a_n x^n - \ldots - a_1 x,$$

$a_0 \epsilon \mathfrak{P}$ donc $a_0 \epsilon \mathfrak{p}$, ce qui est impossible.

Inversement, donnons nous un idéal $\overline{\mathfrak{P}}$ dans $k[x]$, premier, et différent de l'anneau $k[x]$. Nous voulons trouver $\mathfrak{P}$, premier dans $A[x]$, tel que

$$\mathfrak{P} \cap A = \mathfrak{p} \qquad \text{et} \qquad \overline{\mathfrak{P}} = \bar{\lambda}(\mathfrak{P}).$$

L'ensemble $\mathfrak{P}$ des polynomes $F(x)$ qui vérifient

$$\lambda F(x) \epsilon \overline{\mathfrak{P}}$$

est un idéal premier; il satisfait $\mathfrak{P} \cap A = \mathfrak{p}$, car

$$\overline{\mathfrak{P}} \neq k[x];$$

il est différent de $A[x]$ puisque $\mathfrak{p}$ est différent de A. De plus $\bar{\lambda}(\mathfrak{P})$ décrit $\overline{\mathfrak{P}}$, tout polynome de $\overline{\mathfrak{P}}$ pouvant s'écrire

$$\frac{\lambda F(x)}{\overline{w}} \quad \text{avec } F(x) \epsilon A[x].$$

La solution $\mathfrak{P}$ ainsi trouvée est unique; soit $\mathfrak{P}'$ un autre idéal premier tel que

$$\mathfrak{P}' \cap A = \mathfrak{p} \quad \text{et} \quad \bar{\lambda}(\mathfrak{P}') = \overline{\mathfrak{P}}.$$

$F'(x) \epsilon \mathfrak{P}'$ entraine par définition $\dfrac{\lambda F'(x)}{\overline{w}} \epsilon \overline{\mathfrak{P}}$, ou $\lambda F'(x) \epsilon \overline{\mathfrak{P}}$, ce qui prouve

$$\mathfrak{P}' \subseteq \mathfrak{P}.$$

Supposons maintenant $F(x)\epsilon\mathfrak{P}$, donc $\lambda F(x)\epsilon\overline{\mathfrak{P}}$. L'idéal $\overline{\mathfrak{P}}$ est principal dans $k[x]$; on peut prendre sa base sous la forme $\lambda\varphi(x)$, irréductible dans $k[x]$, avec $\varphi(x)\epsilon A[x]$. Il existe donc $P'_0(x)\epsilon\mathfrak{P}'$ et $w'_0\not\equiv 0(\mathfrak{p})$ tels que

$$\overline{\varphi(x)} = \lambda\varphi(x) = \frac{\lambda P'_0(x)}{\overline{w}'_0}.$$

Il vient alors

$$\lambda F(x) = K_1(x)\frac{\lambda P'_0(x)}{\overline{w}'_0} = \frac{\lambda A_1(x)}{\overline{w}}\frac{\lambda P'_0(x)}{\overline{w}'_0}.$$

On en déduit dans $A[x]$:

$$w \,.\, w'_0 \,.\, F(x) = A_1(x)P'_0(x) + \Pi(x)$$

le polynome $\Pi(x)$ étant à coefficients dans $\mathfrak{p}$. Le deuxième membre appartient à $\mathfrak{P}'$, donc le premier. On a

$$w \,.\, w'_0 \not\equiv 0 \ (\mathfrak{p}), \text{ donc } w \,.\, w'_0 \not\equiv 0(\mathfrak{P}'),$$

et, comme $\mathfrak{P}'$ est premier

$$F(x)\epsilon\mathfrak{P}'$$

ce qui démontre

$$\mathfrak{P} \subseteq \mathfrak{P}'.$$

Il en résulte:

$$\mathfrak{P}' = \mathfrak{P}.$$

2. Le théorème de transfert de A à $A[x_1, x_2, \ldots x_n]$. Pour démontrer qu'une propriété de l'anneau A se conserve pour l'anneau $A[x_1, x_2, \ldots, x_n]$ des polynomes à n variables, il suffit de montrer qu'elle s'étend à l'anneau des polynomes à une variable $A[x]$.

Nous allons supposer pour A la

PROPRIÉTÉ 1. *Tout idéal* $\mathfrak{i}$, *différent de l'anneau* A, *admet au moins un diviseur premier* $\mathfrak{p}$, *différent de* A, *tel que pour tout élément* p *de* $\mathfrak{p}$, *il existe au moins un entier* ρ *et un élément* l *associés, vérifiant*

$$p^\rho \,.\, l \equiv 0(\mathfrak{i}) \quad \text{avec} \quad l \not\equiv 0(\mathfrak{p}).$$

Indiquons d'abord deux conséquences de cette propriété.

CONSÉQUENCE 1. *Les éléments* q *tels qu'il existe* l *vérifiant*

$$q \,.\, l \equiv 0(\mathfrak{i}) \quad \text{avec} \quad l \not\equiv 0(\mathfrak{p})$$

forment un idéal primaire $\mathfrak{q}$ *admettant* $\mathfrak{p}$ *pour idéal premier associé.*

On a en effet les trois propriétés suivantes:

(1) $$\mathfrak{i} \subseteq \mathfrak{q} \subseteq \mathfrak{p}$$

(2) $a \,.\, b \equiv 0(\mathfrak{q})$ et $a \not\equiv 0(\mathfrak{q})$ entrainent $b \equiv 0(\mathfrak{p})$.

(3) Si $b \equiv 0(\mathfrak{p})$, il existe un entier ρ tel que $b^\rho \equiv 0(\mathfrak{q})$.

Les deux premières s'établissent immédiatement, et la dernière est une conséquence de la propriété 1. Or ce sont là des conditions caractéristiques

pour que $\mathfrak{q}$ soit primaire, avec $\mathfrak{p}$ pour idéal premier associé. ([2], p. 129, ou [7], §82, p. 33.)

CONSÉQUENCE 2. *L'idéal $\mathfrak{p}$ qui intervient dans la propriété* 1 *est un diviseur premier minimal pour* $\mathfrak{i}$, *c'est à dire*

$$\mathfrak{i} \subseteq \mathfrak{p}' \subseteq \mathfrak{p} \text{ entrainent } \mathfrak{p}' = \mathfrak{p}$$

lorsque $\mathfrak{p}'$ est premier.

En effet, supposons

$$p \equiv 0(\mathfrak{p}) \quad \text{et} \quad p \not\equiv 0(\mathfrak{p}').$$

On a

$$p^\rho . l \equiv 0(\mathfrak{i}) \quad \text{avec} \quad l \not\equiv 0(\mathfrak{p}),$$

d'où

$$p^\rho . l \equiv 0(\mathfrak{p}'),$$

et, puisque $\mathfrak{p}'$ est premier et $p \not\equiv 0(\mathfrak{p}')$,

$$l \equiv 0(\mathfrak{p}'),$$

donc

$$l \equiv 0(\mathfrak{p}),$$

ce qui est contraire à l'hypothèse.

La propriété 1 est satisfaite dans les deux exemples suivants:

EXEMPLE 1. *L'anneau A est l'anneau $K[x]$ des polynomes à une variable à coefficients dans un corps K.*

Quand l'idéal $\mathfrak{i}$ est nul, la solution unique est $\mathfrak{p} = 0$. Quand $\mathfrak{i} \neq 0$, il n'admet qu'un nombre fini de diviseurs premiers qui vérifient tous la propriété 1.

EXEMPLE 2. *L'anneau A est intersection d'un nombre fini d'idéaux primaires.*[2]

Prenons

$$\mathfrak{i} = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_r:$$

décomposition qu'on peut supposer "normée" au sens de Krull ([4], p. 6), c'est à dire dans laquelle aucun des $\mathfrak{q}_j$ $(j = 1, 2, \ldots, r)$ n'est superflu, tandis que les idéaux premiers associés $\mathfrak{p}_j$ sont distincts, et différents de l'anneau A. Parmi ces $\mathfrak{p}_j$, en nombre fini, il existe au moins un idéal minimal $\mathfrak{p} = \mathfrak{p}_1$. Aucun des $\mathfrak{q}_j$ $(j \neq 1)$ n'admet $\mathfrak{p}$ pour idéal premier associé. Pour établir la propriété 1 il suffit alors d'utiliser un raisonnement classique ([7], §83, p. 38).

Soit

$$p \equiv 0(\mathfrak{p}).$$

Il existe un entier ρ tel que

$$p^\rho \equiv 0(\mathfrak{q}_1).$$

$\mathfrak{p}$ étant minimal parmi les $\mathfrak{p}_j$, on peut trouver dans chaque $\mathfrak{p}_j$ $(j \neq 1)$ un élément $l_j \not\equiv 0(\mathfrak{p})$; il vérifie

$$l_j^{\rho_j} \equiv 0(\mathfrak{q}_j).$$

[2] En particulier, moyennant l'axiome du choix, quand A est un anneau noetherien ([7] § 83).

Posons

$$l = \prod_{j \neq 1} l_j^{\rho_j} \not\equiv 0(\mathfrak{p}).$$

Il vient alors

$$\mathfrak{p}^{\rho} \cdot l \equiv 0(\mathfrak{i}) \quad \text{avec} \quad l \not\equiv 0(\mathfrak{p}),$$

ce qui vérifie la propriété 1 dans l'exemple 2. Cette propriété est d'ailleurs vérifiée pour *tout* idéal $\mathfrak{p}'$ diviseur premier minimal pour $\mathfrak{i}$, car $\mathfrak{p}'$ est diviseur d'un $\mathfrak{q}_j$, soit $\mathfrak{q}$, donc d'un $\mathfrak{p}_j$, soit $\mathfrak{p}$; comme il est minimal pour $\mathfrak{i}$, on a $\mathfrak{p}' = \mathfrak{p}$, et $\mathfrak{p}$ est nécessairement minimal parmi les $\mathfrak{p}_j$. On termine alors comme plus haut.

Nous avons en vue le transfert de la propriété 1, supposée valable pour l'anneau A, à l'anneau $A[x]$ des polynomes à une variable. Soit $\mathfrak{J}$ un idéal dans $A[x]$, différent de $\mathfrak{R} = A[x]$. Considérons, comme au paragraphe 1, la projection

$$\mathfrak{i} = \mathfrak{J} \cap A$$

de $\mathfrak{J}$ dans l'anneau A. C'est un idéal $\mathfrak{i}$, différent de A. Par hypothèse l'idéal $\mathfrak{i}$ possède un diviseur premier $\mathfrak{p}$ qui vérifie la propriété 1. La recherche d'un diviseur premier $\mathfrak{P}$ pour $\mathfrak{J}$ est alors liée au problème suivant:

Problème d'extension. $\mathfrak{i}$ *et* $\mathfrak{p}$ *étant des idéaux dans* A *qui satisfont la propriété* 1, $\mathfrak{J}$ *un idéal dans* $A[x]$ *qui se projette suivant* $\mathfrak{i}$ *dans* A, *trouver un diviseur premier* $\mathfrak{P}$ *de* $\mathfrak{J}$, *qui se projette suivant* $\mathfrak{p}$ *dans* A.

Les données sont $\mathfrak{J}$, $\mathfrak{i} = \mathfrak{J} \cap A$ et $\mathfrak{p}$ avec la propriété 1 pour $\mathfrak{i}$. Il faut trouver un diviseur premier $\mathfrak{P}$ de $\mathfrak{J}$ tel que $\mathfrak{P} \cap A = \mathfrak{p}$. La solution de ce problème n'est pas sans analogie avec celle de l'extension d'une spécialisation, donnée par A. Weil ([8], p. 30).

Considérons comme au paragraphe 1, l'idéal $\overline{\mathfrak{J}} = \overline{\lambda}(\mathfrak{J})$. *L'idéal* $\overline{\mathfrak{J}} = \overline{\lambda}(\mathfrak{J})$ *n'est pas l'idéal unité dans* $k[x]$. S'il était l'idéal unité, une constante non nulle serait dans $\overline{\mathfrak{J}}$ soit

$$\frac{\bar{u}'}{\bar{w}'} = \frac{\lambda F_0(x)}{\bar{w}}$$

avec

$$F_0(x) = a_n x^n + \ldots + a_1 x + a_0 \in \mathfrak{J}$$

et, puisque $\dfrac{\lambda F_0(x)}{\bar{w}}$ est une constante non nulle de $k[x]$,

$$\bar{a}_n = 0, \ldots, \bar{a}_1 = 0; \ \bar{a}_0 \neq 0$$

ou

$$a_n \equiv \ldots \equiv a_1 \equiv 0 \ (\mathfrak{p}); \ a_0 \not\equiv 0 \ (\mathfrak{p}).$$

D'après la propriété 1 appliquée à $\mathfrak{i}$ et $\mathfrak{p}$, on peut trouver des entiers ρ_j et des éléments $l_j \not\equiv 0(\mathfrak{p})$, tels que

$$a_j^{\rho_j} \cdot l_j \equiv 0(\mathfrak{i}); \ l_j \not\equiv 0(\mathfrak{p}); \ j = n, \ldots, 1.$$

Posons

$$\rho = 1 + \sum_{j=1}^{j=n} (\rho_j - 1); \; l = \prod_j l_j \not\equiv 0(\mathfrak{p}).$$

De la relation

$$a_0 = F_0(x) - a_n x^n - \ldots - a_1 x$$

on tire

$$a_0^\rho \,.\, l \equiv 0(\mathfrak{i}), \; l \not\equiv 0(\mathfrak{p}),$$

ce qui prouve

$$a_0 \epsilon \mathfrak{p}$$

contrairement à l'hypothèse.

Il ne reste plus pour l'idéal que deux possibilités:

1° cas. *L'idéal* $\overline{\mathfrak{F}}$ *est l'idéal nul*, c'est à dire pour tout

$$F(x) = a_n x^n + \ldots + a_0 \epsilon \mathfrak{F}$$

on a

$$\lambda F(x) = 0, \quad \text{ou} \quad \bar{a}_n = \ldots = \bar{a}_0 = 0$$

ce qui signifie

$$a_n \equiv \ldots \equiv a_0 \equiv 0(\mathfrak{p}).$$

L'idéal $\mathfrak{F}$ est donc contenu dans l'idéal Π des polynomes de $A[x]$ qui ont leurs coefficients dans $\mathfrak{p}$.

Cet idéal Π, formé par les polynomes $F(x)$ tels que

$$\lambda F(x) = 0$$

est d'après le lemme 1, l'idéal premier unique, différent de $A[x]$, vérifiant

$$\Pi \cap A = \mathfrak{p} \quad \text{et} \quad \bar{\lambda}(\Pi) = 0.$$

C'est donc un diviseur premier de $\mathfrak{F}$, solution du problème d'extension. D'ailleurs, toute solution du problème d'extension contient $\mathfrak{p}$ donc Π.

L'idéal Π *est donc la solution unique minimale du problème d'extension.*

2° cas. *L'idéal* $\overline{\mathfrak{F}}$ *n'est pas l'idéal nul.* Soit $\mathfrak{P}$ une solution du problème d'extension. L'idéal $\overline{\mathfrak{P}} = \bar{\lambda}(\mathfrak{P})$ est un diviseur premier de $\overline{\mathfrak{F}} \neq 0$. Or ces diviseurs sont en nombre fini; ils ont pour base l'un des facteurs irréductibles $\overline{\varphi_j(x)}$ de la décomposition de la base $\overline{\varphi(x)}$ de $\overline{\mathfrak{F}}$ dans $k[x]$. $(j = 1, 2, \ldots, r)$. Chacun d'eux, $\overline{\mathfrak{P}}_j$, ne correspond d'après le lemme 1 qu'à un seul idéal premier $\mathfrak{P}_j$ tel que

$$\mathfrak{P}_j \cap A = \mathfrak{p} \quad \text{et} \quad \bar{\lambda}(\mathfrak{P}_j) = \overline{\mathfrak{P}}_j.$$

$\mathfrak{P}_j$ est différent de $\mathfrak{R} = A[x]$, et il est constitué par tous les polynomes $F(x)$ tels que

$$\lambda F(x) \epsilon \overline{\mathfrak{P}}_j.$$

C'est aussi un diviseur premier de $\mathfrak{F}$, car pour tout $F(x) \epsilon \mathfrak{F}$ on a par définition

$$\lambda F(x) \epsilon \overline{\mathfrak{F}}, \text{ donc } \lambda F(x) \epsilon \overline{\mathfrak{P}}_j.$$

Le problème d'extension admet, dans ce deuxième cas, r *solutions* constituées par les r idéaux $\mathfrak{P}_j$.

En résumé, le problème d' extension peut présenter deux cas:

1° cas. *Il existe une solution* Π *et toute solution est un diviseur de* Π.

2° cas. *Il n'y a qu'un nombre fini de solutions constituées par les* r *idéaux* $\mathfrak{P}_j$.

Nous allons voir que la solution unique minimale Π du premier cas, et chacune des solutions $\mathfrak{P}_j$ du 2° cas vérifient la propriété 1 pour l'idéal $\mathfrak{J}$ dans $A[x]$.

Dans le 1° *cas*: Π est un diviseur premier de $\mathfrak{J}$ distinct de l'idéal unité $\mathfrak{R} = A[x]$.

Soit

$$F(x) = a_n x^n + \ldots + a_0 \equiv 0(\Pi)$$

On a donc

$$a_n \equiv \ldots \equiv a_0 \equiv 0(\mathfrak{p})$$

d'où, d'après la propriété 1 appliquée à i et $\mathfrak{p}$,

$$a_j^{\rho_j} \equiv 0(\mathfrak{i}), \quad l_j \not\equiv 0(\mathfrak{p}),$$

ce qui s'écrit encore

$$a_j^{\rho_j} l_j \equiv 0(\mathfrak{J}), \quad l_j \not\equiv 0(\Pi).$$

En posant

$$\rho = 1 + \sum_{j=n}^{j=0} (\rho_j - 1), \quad l = \prod_j l_j,$$

on obtient

$$F^{\rho} . l \equiv 0(\mathfrak{J}), \quad l \not\equiv 0(\Pi).$$

La propriété 1 est bien vérifiée pour $\mathfrak{J}$ et son diviseur premier Π.

Dans le 2° *cas*: Soit $P(x) \epsilon \mathfrak{P}_j = \mathfrak{P}$.

Opérons d'abord dans $k[x]$. Il existe un entier m tel que

$$\overline{P}^m . \overline{L}_0 \equiv 0(\overline{\mathfrak{J}}), \quad \overline{L}_0 \not\equiv 0(\overline{\mathfrak{P}})$$

d'après la propriété 1 valable dans $k[x]$ (exemple 1). On en déduit dans $A[x]$:

$$wL_0 P^m = F(x) + \Pi(x)$$
$$wL_0 \not\equiv 0(\mathfrak{P}), \quad F(x) \epsilon \mathfrak{J}, \quad \Pi(x) \epsilon \Pi.$$

On a vu dans le 1° cas qu'il existe un entier σ et un élément l tels que

$$\Pi(x)^{\sigma} . l \equiv 0 \ (\mathfrak{J}), \quad l \not\equiv 0(\mathfrak{p}).$$

On en tire, en multipliant par l l'expression $w.L_0 P^m$ élevée à la puissance σ,

$$P^{m\sigma} . L \equiv 0(\mathfrak{J}), \quad L \not\equiv 0(\mathfrak{P}),$$

soit

$$P^{\rho} L \equiv 0(\mathfrak{J}), \quad L \not\equiv 0 \ (\mathfrak{P}).$$

Ainsi se trouve démontré le théorème de transfert pour la propriété 1:

THÉORÈME 1. *Quand la propriété* 1 *est valable pour l'anneau* A, *elle s'étend à l'anneau* $A[x_1, x_2, \ldots, x_n]$ *des polynomes à* n *variables.*

En particulier, d'après l'exemple 1,

THÉORÈME 2. *Tout idéal* $\mathfrak{J}$ *dans l'anneau* $K[x_1, \ldots, x_n]$ *des polynomes à* n *variables à coefficients dans un corps* K *admet au moins un diviseur premier*

$\mathfrak{P}$, différent de l'idéal unité, et tel que pour tout polynome P de $\mathfrak{P}$, on puisse trouver un entier ρ et un polynome L vérifiant

$$P^{\rho}.L \equiv 0(\mathfrak{J}),\quad L \not\equiv 0(\mathfrak{P}).$$

3. Le théorème des zéros de Hilbert. K étant un corps quelconque, le théorème des zéros de Hilbert s'énonce ainsi:

THÉORÈME 3. *Quand un polynome $f \in K[x_1, \ldots, x_n]$ s'annule pour tous les zéros algébriques d'un idéal $\mathfrak{J}$, il existe un entier ρ tel que $f^{\rho} \in \mathfrak{J}$.*

La méthode de A. Rabinowitsch [6] exposée par B. L. van der Waerden ([7], §75, p. 11) ramène la démonstration de ce théorème au suivant:

THÉORÈME 4. *Dans l'anneau des polynomes $K[x_1, \ldots, x_n]$, un idéal qui n'a aucun zéro algébrique sur K est nécessairement l'idéal unité.*

Ou encore, un idéal $\mathfrak{J} \neq \mathfrak{R}$ possède toujours au moins un zéro algébrique sur K.

Pour établir le théorème 4, il suffit de le démontrer pour le diviseur premier $\mathfrak{P} \neq \mathfrak{R}$ dont l'existence est assurée par le théorème 2. Nous allons démontrer le théorème suivant, plus précis:

THÉORÈME 5. *Un idéal premier $\mathfrak{P}$ dans $K[x_1, \ldots, x_n]$ qui n'a aucun zéro algébrique sur K, est nécessairement l'idéal unité. De plus, un polynome f qui s'annule pour tout zéro algébrique d'un idéal premier $\mathfrak{P}$ appartient nécessairement à $\mathfrak{P}$.*

Autrement dit, il existe toujours un zéro algébrique sur K d'un idéal premier $\mathfrak{P}$ différent de $\mathfrak{R}$. De plus, si

$$f(x_1, x_2, \ldots, x_n) \not\equiv 0(\mathfrak{P})$$

on peut choisir ce zéro $(x'_1, x'_2, \ldots, x'_n)$ de façon que

$$f(x'_1, x'_2, \ldots, x'_n) \neq 0.$$

C'est sous cette forme que nous allons établir le théorème, par induction sur le nombre des variables. Nous n'aurons besoin pour cela que des remarques préliminaires du paragraphe 1.

Le cas $n = 1$ est bien connu. Supposons le théorème établi dans l'anneau $A = K[x_1, \ldots, x_{n-1}]$ et démontrons le dans l'anneau $A[x_n] = A[x] = K[x_1, \ldots, x_n]$. Soit

$$\mathfrak{p} = \mathfrak{P} \cap A \neq A,$$

la projection de $\mathfrak{P}$ dans A. D'après le lemme 1, l'idéal $\mathfrak{P}$ est complètement déterminé par $\mathfrak{p}$ et par

$$\overline{\mathfrak{P}} = \bar{\lambda}(\mathfrak{P}).$$

1°) $\overline{\mathfrak{P}} = 0$. L'idéal $\mathfrak{P}$ coincide avec l'idéal Π des polynomes

$$F = a_n x^n + \ldots + a_0$$

à coefficients $a_n, \ldots, a_0$ dans $\mathfrak{p}$. Tout zéro algébrique de $\mathfrak{p}$ est alors zéro de $\mathfrak{P}$, quel que soit x. De plus, si

$$f = a_m x^m + \ldots + a_d x^d + \ldots + a_0 \not\equiv 0(\mathfrak{P})$$

l'un de ses coefficients n'est pas dans $\mathfrak{p}$; soit a_d le premier coefficient non dans $\mathfrak{p}$.
Quand $d = 0$, on a $a_0 \not\equiv 0(\mathfrak{p})$, $a_j \equiv 0(\mathfrak{p})$. $(j > 0)$
On peut alors trouver un zéro algébrique de $\mathfrak{p}$, soit $(x'_1, x'_2, \ldots, x'_{n-1})$, tel que

$$a_0(x'_1, \ldots, x'_{n-1}) \not\equiv 0.$$

Alors $(x'_1, x'_2, \ldots, x'_{n-1}, x'_n)$ est un zéro algébrique de $\mathfrak{P}$ dès que x'_n est algébrique sur K, et on a pour ce zéro:

$$f(x'_1, \ldots, x'_n) \not\equiv 0.$$

Quand $d > 0$, on choisit un zéro algébrique $(x'_1, \ldots, x'_{n-1})$ de $\mathfrak{p}$ tel que

$$a'_d = a_d(x'_1, \ldots, x_{n-1}) \not\equiv 0,$$

puis pour x_n un nombre algébrique sur K, distinct des racines de l'équation

$$a'_d x^d + \ldots + a'_0 = 0.$$

Cela est toujours possible car il y a toujours une infinité de nombres algébriques sur K. On obtient un zéro $(x'_1, \ldots, x'_{n-1}, x'_n)$ de $\mathfrak{P}$, algébrique sur K, tel que

$$f(x'_1, \ldots, x'_n) \not\equiv 0.$$

2°) $\mathfrak{P} \not\equiv 0$. On peut prendre la base de $\overline{\mathfrak{P}}$ sous la forme $\lambda\Psi(x) \not\equiv 0$ avec

$$\Psi(x) = a_d x^d + \ldots + a_0 \in \mathfrak{P}$$

et

$$a_d \not\equiv 0(\mathfrak{p}),\ d > 0.$$

Tout polynome $F(x) \in A[x]$ peut être divisé par $\Psi(x)$ pour donner dans $A[x]$:

$$a_d^\rho F = Q \,.\, \Psi + R, \qquad d^0 R < d.$$

Si $F(x)$ est pris dans $\mathfrak{P}$, $\lambda R \in \mathfrak{P}$ car $\lambda\Psi \in \mathfrak{P}$.
On obtient alors

$$\lambda R = 0 \quad \text{ou} \quad R \in \Pi$$

sans quoi le degré de λR serait inférieur au degré d de la base $\lambda\Psi$ de $\overline{\mathfrak{P}}$. Donc, pour tout polynome $P \in \mathfrak{P}$ il existe a_d et $\Psi(x)$, fixes, et ρ, $Q(x)$, $\Pi(x)$ variables avec P, tels que

$$(1) \qquad a_d^\rho P(x) = Q(x) \,.\, \Psi(x) + \Pi(x),\ a_d \not\equiv 0(\mathfrak{p}).$$

Soit

$$f(x_1, x_2, \ldots, x_n) \not\equiv 0(\mathfrak{P}).$$

Donc, dans $k[x]$, le polynome $\bar{f} = \lambda f$ est non multiple de $\overline{\Psi} = \lambda\Psi$, par suite premier avec lui. On peut alors écrire dans $k[x]$

$$1 = k_1(x) \,.\, \bar{f} + k_2(x) \,.\, \overline{\Psi},$$

ce qui donne dans $A[x] = K[x_1, \ldots, x_n]$

$$(2) \qquad w = U \,.\, f + V \,.\, \Psi + \Pi(x)$$

où w est dans A mais non dans $\mathfrak{p}$, tandis que $\Pi(x)$ est à coefficients dans $\mathfrak{p}$.

On peut trouver un zéro de $\mathfrak{p}$, algébrique sur K, soit $x'_1, \ldots, x'_{n-1}$, qui n'annule pas $w \,.\, a_d \not\equiv 0(\mathfrak{p})$. Il est donc tel que

$$w' \not\equiv 0 \quad \text{et} \quad a'_d \not\equiv 0.$$

Prenons pour $x_n = x$ une racine de

$$a'_d x^d + \ldots + a'_0 = 0.$$

On obtient, d'après (1), un zéro $(x'_1, \ldots, x'_n)$ de $\mathfrak{P}$, algébrique sur K, et ce zéro, d'après (2), ne peut annuler f:

$$f(x'_1, \ldots, x'_n) \not\equiv 0.$$

Le théorème 5 est démontré, ainsi que le théorème des zéros de Hilbert.

Des démonstrations du théorème des zéros de Hilbert se trouvent aussi dans O. Zariski [9], qui suppose connue l'existence d'un idéal maximal, diviseur d'un idéal donné; cette affirmation nécessite l'axiome du choix; (cf. N. Bourbaki, traité de Mathématiques, *Algèbre*, chap. 1, §8, n°. 7). La démonstration exposée dans le traité de van der Waerden ([7], p. 10), utilise la théorie de l'élimination. Il existe aussi une démonstration élémentaire de R. Brauer [1].

4. Décomposition d'un idéal dans $K[x_1, \ldots, x_n]$ en idéaux primaires. L'hypothèse de la base finie dans un anneau A entraine comme on le sait la propriété des "chaines de diviseurs, "([2] p. 134, ou [7] p. 26), qui exprime qu'une suite non décroissante d'idéaux

$$\mathfrak{J}_1 \subseteq \mathfrak{J}_2 \subseteq \ldots \subseteq \mathfrak{J}_k \subseteq \mathfrak{J}_{k+1} \subseteq \ldots$$

est stationnaire à partir d'un certain rang k.

$$\mathfrak{J}_k = \mathfrak{J}_{k+j} \text{ pour tout } j > 0.$$

La propriété I, au §2, a pour conséquence (conséquence I): les éléments Q tels qu'il existe un élément L vérifiant

$$Q \,.\, L \equiv 0(\mathfrak{J}), \quad L \not\equiv 0(\mathfrak{P})$$

forment un idéal primaire $\mathfrak{Q}$ admettant $\mathfrak{P}$ pour idéal premier associé.

On a évidemment

$$\mathfrak{J} \subseteq \mathfrak{Q} \subseteq \mathfrak{P}.$$

Il est possible de préciser dans *l'hypothèse de la base finie.*

LEMME 2. *L'idéal $\mathfrak{J}$ est alors l'intersection de l'idéal primaire $\mathfrak{Q}$ et d'un idéal,*

$$\mathfrak{J} + (M)$$

qui n'admet plus $\mathfrak{P}$ comme diviseur premier.

(M) désigne l'idéal principal de base M, et $\mathfrak{J}+(M)$ l'idéal ayant pour base l'ensemble constitué par la base de $\mathfrak{J}$ et par M.

Soit

$$\mathfrak{Q} = (Q_1, \ldots, Q_s).$$

Il existe pour chaque $Q_j (j = 1, 2, \ldots, s)$, un élément $L_j \not\equiv (\mathfrak{P})$ tel que

$$Q_j L_j \equiv 0(\mathfrak{J}).$$

On a donc, en posant $L = \prod_j L_j$,

$$\mathfrak{Q} \cdot L = 0(\mathfrak{J}) \text{ avec } L \not\equiv 0(\mathfrak{P}).$$

D'autre part, d'après un raisonnement connu ([7] §82, p. 36) la propriété des chaines de diviseurs entraine l'existence d'un entier K tel que[3]

$$\mathfrak{J} : L^k = \mathfrak{J} : L^{k+1}.$$

Posons $L^k = M \not\equiv 0(\mathfrak{P})$; nous allons montrer que

$$\mathfrak{J} = \mathfrak{Q} \cap (\mathfrak{J} + (M)).$$

$\mathfrak{J}$ est évidemment inclus dans l'intersection de $\mathfrak{Q}$ et de $\mathfrak{J} + (M)$. Réciproquement, soit c un élément de cette intersection:

$$c = a + bL^k \text{ avec } a \epsilon \mathfrak{J}.$$

c appartenant à $\mathfrak{Q}$, on a

$$cL \epsilon \mathfrak{J} \quad \text{d'où} \quad bL^{k+1} \epsilon \mathfrak{J}.$$

Par suite

$$b \epsilon \mathfrak{J} : L^{k+1} = \mathfrak{J} : L^k.$$

Donc

$$bL^k \epsilon \mathfrak{J}$$

et

$$c \epsilon \mathfrak{J}.$$

Le lemme 2 s'applique naturellement à l'anneau $K[x_1, \ldots, x_n]$, d'après le théorème de la base finie de Hilbert, et d'après le théorème 2.

Il suffit pour établir la décomposition en idéaux primaires de tout idéal dans $K[x_1, \ldots, x_n]$. En effet, si l'idéal

$$\mathfrak{J}_1 = \mathfrak{J} + (M) \supset \mathfrak{J}$$

n'est pas l'idéal unité, il admet d'après le théorème 2 un diviseur premier $\mathfrak{P}_1 \not\equiv \mathfrak{R}$ auquel le lemme 2 s'applique. On forme

$$\mathfrak{J}_2 = \mathfrak{J}_1 + (M_1) \supset \mathfrak{J}_1.$$

L'opération prend fin après un nombre fini k d'opérations; car en la poursuivant indéfiniment on formerait une chaine infinie d'idéaux

$$\mathfrak{J} \subset \mathfrak{J}_1 \subset \mathfrak{J}_2 \subset \ldots \subset \mathfrak{J}_k \subset \ldots$$

ce qui est contraire à la propriété des chaines de diviseurs. L'idéal $\mathfrak{J}_{k+1}$ est donc l'idéal unité, et on a

$$\mathfrak{J} = \mathfrak{Q} \cap \mathfrak{Q}_1 \cap \ldots \cap \mathfrak{Q}_k.$$

La démonstration suivante, plus longue, il est vrai, révèle mieux, à ce qu'il me semble, la structure des idéaux de polynomes à coefficients dans un corps K. Elle parait plus adaptée à l'étude des variétés algébriques.

D'autre part elle dispense de l'axiome du choix, sous la forme restreinte utilisée dans la démonstration précédente comme dans la démonstration de E. Noether. ([7] §83.)

[3] On forme la chaine de diviseurs:

$$\mathfrak{Q} = \mathfrak{J} : L \subseteq \mathfrak{J} : L^2 \subseteq \ldots \subseteq \mathfrak{J} : L^k \subseteq \mathfrak{J} : L^{k+1} \subseteq \ldots$$

Une notion importante dans $K[x_1, \ldots, x_n]$, est celle de dimension d'un idéal premier $\mathfrak{P}$; soient $\xi_1, \ldots, \xi_n$ les classes de $x_1, \ldots, x_n$ modulo $\mathfrak{P}$.

L'anneau $K[x_2, \ldots, x_n]/\mathfrak{P}$ est un domaine d'intégrité dont le corps des quotients est

$$K(\xi_1 \ldots, \xi_n).$$

La dimension de $\mathfrak{P}$ est alors la dimension sur K du corps $K(\xi_1, \ldots, \xi_n)$. C'est le nombre maximum d'éléments algébriquement indépendants sur K parmi les $\xi_j (j = 1, 2, \ldots, n)$. Rappelons la propriété suivante concernant les dimensions: ([7], §90, p. 63 ou [8] chap. II, th. 3, p. 28).

LEMME 3. *Soit $\mathfrak{P}$ un idéal premier dans $\mathfrak{R} = K[x_1, \ldots, x_n]$, différent de $\mathfrak{R}$, et de dimension d. Soit $\mathfrak{P}'$ un diviseur premier de $\mathfrak{P}$, différent de $\mathfrak{R}$, de dimension d'. On a*

$$d' \leqslant d$$

et l'égalité n'a lieu que si $\mathfrak{P}' = \mathfrak{P}$.

Maintenant, la préparation du théorème suivant est suffisante:

THÉORÈME 6. *Tout idéal $\mathfrak{J}$ dans l'anneau des polynomes $\mathfrak{R} = K[x_1, \ldots, x_n]$ est intersection d'un nombre fini d'idéaux primaires.*

Si $\mathfrak{J}$ remplit $\mathfrak{R}$, il est premier, donc primaire. Nous pouvons donc supposer $\mathfrak{J} \neq \mathfrak{R}$. Le raisonnement s'effectue par induction sur le nombre des variables.

Le cas $n = 1$ est bien connu. Supposons alors le théorème valable dans $K[x_1, \ldots, x_{n-1}] = A$ et considérons la projection de $\mathfrak{J}$ dans A:

$$\mathfrak{i} = \mathfrak{J} \cap A.$$

C'est une intersection d'idéaux primaires

$$\mathfrak{i} = \mathfrak{q}_1 \cap \mathfrak{q}_2 \cap \ldots \cap \mathfrak{q}_t$$

chaque $\mathfrak{q}_j$ ayant pour idéal premier associé $\mathfrak{p}_j$. Les $\mathfrak{p}_j$ de dimension maximum d $(0 \leqslant d \leqslant n-1)$ sont nécessairement minimaux pour $\mathfrak{i}$, car si $\mathfrak{p}$ est l'un de ces idéaux, un diviseur premier $\mathfrak{p}'$ tel que

$$\mathfrak{i} \subseteq \mathfrak{p}' \subseteq \mathfrak{p}$$

serait diviseur d'un $\mathfrak{p}_j$ (cf. exemple 2); sa dimension, d'après le lemme 3, serait supérieure ou égale à celle de $\mathfrak{p}$ et inférieure ou égale à celle de $\mathfrak{p}_j$; comme la dimension de $\mathfrak{p}_j$ ne peut dépasser celle de $\mathfrak{p}$, l'égalité est obligatoire, ce qui entraine d'après le lemme 3

$$\mathfrak{p}' = \mathfrak{p}.$$

$\mathfrak{p}$ étant minimal pour $\mathfrak{i}$, nous avons vu (exemple 2) que ces deux idéaux vérifient la propriété 1. Le problème d'extension concernant $\mathfrak{i}$, $\mathfrak{p}$ et $\mathfrak{J}$ se traite alors avec succès. Deux cas peuvent se présenter, d'après l'étude faite au paragraphe 2.

1°) *Il n'y a qu'un nombre fini de solutions $\mathfrak{P}_1, \ldots, \mathfrak{P}_j$ qui se projettent suivant $\mathfrak{p}$.*

(C'etait le deuxième cas envisagé au paragraphe 2.)

Prenons l'idéal $\mathfrak{P}_1$ et soit $\mathfrak{Q}_1$ l'idéal primaire correspondant. On a (lemme 2):

$$\mathfrak{J} = \mathfrak{Q}_1 \cap (\mathfrak{J} + (M_1);$$

aucun des $\mathfrak{P}_j$ $(j = 2, \ldots, r)$ n'est diviseur de $\mathfrak{Q}_1$, car ils sont tous minimaux pour $\mathfrak{J}$, donc pour $\mathfrak{Q}_1$ ou pour $\mathfrak{J} + (M_1) = \mathfrak{J}_1$ et le seul diviseur premier minimal de $\mathfrak{Q}_1$ est $\mathfrak{P}_1$. Chacun des $\mathfrak{P}_j (j \geqslant 2)$ doit donc être diviseur premier minimal pour l'idéal $\mathfrak{J}_1$. D'ailleurs, tout diviseur premier de $\mathfrak{J}_1$ qui se projette suivant $\mathfrak{p}$ est nécessairement diviseur premier de $\mathfrak{J}$ ayant $\mathfrak{p}$ pour projection dans A; il coïncide avec l'un des $\mathfrak{P}_j (j \geqslant 2)$. Posons

$$\mathfrak{i}_1 = \mathfrak{J}_1 \cap A.$$

Des relations

$$\mathfrak{J} \subseteq \mathfrak{J}_1 \subseteq \mathfrak{P}_2$$

on tire en projetant dans $A = K[x_1, \ldots, x_{n-1}]$:

$$\mathfrak{i} \subseteq \mathfrak{i}_1 \subseteq \mathfrak{p}.$$

$\mathfrak{p}$ est donc diviseur premier de $\mathfrak{i}_1$, minimal pour $\mathfrak{i}_1$ puisqu'il l'est pour $\mathfrak{i}$. Il vérifie pour $\mathfrak{i}_1$ la propriété 1 (exemple 2). Le problème d'extension traité pour $\mathfrak{i}_1$, $\mathfrak{p}$ et $\mathfrak{J}_1$ donne nécessairement pour solutions $\mathfrak{P}_2, \ldots, \mathfrak{P}_r$. Considérons $\mathfrak{P}_2$ et l'idéal primaire associé $\mathfrak{q}'_2$. Le lemme 2 donne

$$\mathfrak{J}_1 = \mathfrak{Q}'_2 \cap (\mathfrak{J}_1 + (M_2)), \qquad M_2 \not\equiv 0(\mathfrak{P}_2).$$

D'où en posant

$$\mathfrak{J}_2 = \mathfrak{J}_1 + (M_2)$$

il vient

$$\mathfrak{J} = \mathfrak{Q}_1 \cap \mathfrak{Q}'_2 \cap \mathfrak{J}_2,$$

et comme

$$\mathfrak{J} \subseteq \mathfrak{Q}_2 \subseteq \mathfrak{Q}'_2$$

on a aussi

$$\mathfrak{J} = \mathfrak{Q}_1 \cap \mathfrak{Q}_2 \cap \mathfrak{J}_2$$

avec

$$\mathfrak{J}_2 = \mathfrak{J} + (M_1) + (M_2).$$

On démontre, comme on l'a fait pour $\mathfrak{J}_1$, que l'idéal $\mathfrak{J}_2$ admet comme diviseurs premiers qui se projettent suivant $\mathfrak{p}$ dans A, les idéaux, $\mathfrak{P}_3, \ldots, \mathfrak{P}_r$, et qu'il n'admet que ceux là . On arrive ainsi de proche en proche à la décomposition

$$\mathfrak{J} = \mathfrak{Q}_1 \cap \mathfrak{Q}_2 \cap \ldots \cap \mathfrak{Q}_r \cap \mathfrak{A}$$

avec

$$\mathfrak{A} = \mathfrak{J} + (M_1) + \ldots + (M_r) \qquad M_j \not\equiv 0(\mathfrak{P}_j)$$

et

$$\mathfrak{a} = \mathfrak{A} \cap A \not\equiv 0(\mathfrak{p})$$

car si $\mathfrak{a} \subseteq \mathfrak{p}$, les solutions du problème d'extension pour $\mathfrak{a}$, $\mathfrak{p}$ et $\mathfrak{A}$ seraient à choisir parmi les $\mathfrak{P}_j$, dont aucun ne convient.

2°) *La solution minimale du problème d'extension est l'idéal* Π *des polynomes* $F(x)$ *dont les coefficients dans* A *appartiennent à* $\mathfrak{p}$.

Le lemme 2 donne

$$\mathfrak{J} = \mathfrak{Q} \cap \mathfrak{J}',$$

$\mathfrak{Q}$ étant l'idéal primaire correspondant à Π, avec

$$\mathfrak{J}' = \mathfrak{J} + (M), \qquad M \not\equiv 0(\Pi).$$

La projection

$$\mathfrak{i}' = \mathfrak{J}' \cap A$$

de $\mathfrak{J}'$ dans A peut encore admettre $\mathfrak{p}$ comme diviseur premier, nécessairement minimal. Le problème d'extension traité pour $\mathfrak{i}'$, $\mathfrak{J}'$ et $\mathfrak{p}$ donne alors une ou plusieurs solutions minimales diviseurs premiers de Π, mais il ne peut rentrer dans le même type que le problème concernant $\mathfrak{i}$, $\mathfrak{J}$ et $\mathfrak{p}$ car sa solution minimale serait Π ce qui est impossible puisque $M \not\equiv 0(\Pi)$, et par suite $\mathfrak{J}' \not\equiv 0(\Pi)$. Il rentre alors dans le type traité précédemment, ce qui permet d'écrire

$$\mathfrak{J} = \mathfrak{Q} \cap \mathfrak{Q}_1 \cap \ldots \cap \mathfrak{Q}_r \cap \mathfrak{A}.$$

Les idéaux $\mathfrak{Q}$, $\mathfrak{Q}_1, \ldots, \mathfrak{Q}_r$ sont primaires et leurs idéaux premiers se projettent dans A suivant $\mathfrak{p}$, l'idéal $\mathfrak{A}$ se projette dans A suivant un idéal $\mathfrak{a}$ n'admettant plus $\mathfrak{p}$ pour diviseur premier.[4] Les diviseurs premiers de $\mathfrak{a}$ ayant la dimension d sont nécessairement minimaux pour $\mathfrak{a}$ car se sont des diviseurs premiers minimaux de $\mathfrak{i}$. Ils sont en nombre fini; on les épuise par la méthode précédente en raisonnant sur $\mathfrak{A}$ au lieu de $\mathfrak{J}$. On met ainsi $\mathfrak{J}$ sous la forme:

$$\mathfrak{J} = \mathfrak{J}_d \cap \mathfrak{F}$$

où $\mathfrak{J}_d$ est l'intersection d'idéaux primaires dont les projections dans A appartiennent à des idéaux premiers de dimension d, et $\mathfrak{F}$ un idéal dont la projection dans A n'admet aucun diviseur premier de dimension supérieure ou égale à d.

Dans la deuxième étape, on se débarrasse en raisonnant sur $\mathfrak{F}$, des diviseurs premiers de sa projection dont la dimension est $d - 1$. Et ainsi de suite, après chaque étape la dimension des diviseurs premiers de la projection descend. On arrive, si l'idéal unité n'a pas été rencontré en chemin, à une projection dont les diviseurs premiers sont de dimension zéro, donc tous minimaux. On s'en débarrasse de la même façon; il reste un idéal sans diviseur premier minimal, c'est à dire l'idéal unité. Par suite

$$\mathfrak{J} = \mathfrak{J}_d \cap \mathfrak{J}_{d-1} \cap \ldots \cap \mathfrak{J}_0,$$

l'idéal $\mathfrak{J}_h$ rassemblant l'intersection des idéaux primaires en nombre fini, dont les projections dans A ont des diviseurs premiers associés de dimension h.

Les théorèmes d'unicité sur la décomposition s'établissent ensuite par la méthode classique, sans utiliser l'axiome du choix (van der Waerden [7], §84, pp. 39-43).

[4] Les diviseurs premiers minimaux de $\mathfrak{a}$ comprennent en particulier les diviseurs premiers minimaux de $\mathfrak{i}$ distincts de $\mathfrak{p}$, car si $\mathfrak{p}'$ est l'un de ces diviseurs il donne naissance à des diviseurs premiers minimaux de $\mathfrak{J}$ qui sont diviseurs premiers minimaux de $\mathfrak{Q}$ ou $\mathfrak{Q}_j$ ($j = 1, 2, \ldots, r$) ou $\mathfrak{A}$. Ils ne peuvent l'être pour $\mathfrak{Q}$ ou $\mathfrak{Q}_j$ car ils coïncideraient avec π ou $\mathfrak{Q}_j$. Leur projection dans A serait $\mathfrak{p}$ au lieu de $\mathfrak{p}'$.

References

[1] R. Brauer, "A Note on Hilbert's Nullstellensatz," *Bull. Amer. Math. Soc.*, vol. 54 (1948), 894-896.
[2] P. Dubreil, Algèbre, tome I (Gauthier-Villars, 1946).
[3] D. Hilbert, "Über die Theorie der algebraischen Formen," *Math. Ann.*, vol. 36 (1890), 473.
[4] W. Krull, *Idealtheorie* (Ergebnisse der Math., IV, 3, 1935).
[5] E. Lasker, "Zur Theorie der Moduln und Ideale," *Math. Ann.*, vol. 60 (1905), 20, 116.
[6] A. Rabinowitsch, *Math. Ann.*, vol. 102 (1929), 518.
[7] B. L. van der Waerden, *Moderne Algebra*, tome II (Grundl. der Math. Wissensch., vol. 34. Berlin, 1931).
[8] A. Weil, *Foundations of Algebraic Geometry* (Amer. Math. Soc. Colloq. Publications, vol. 29, 1946).
[9] O. Zariski, "A New Proof of Hilbert's Nullstellensatz," *Bull. Amer. Math. Soc.*, vol. 53 (1947), 362-368.

Tours, France

AN ELEMENTARY PROOF OF THE PRIME-NUMBER THEOREM FOR ARITHMETIC PROGRESSIONS

ATLE SELBERG

1. Introduction. In this paper we shall give an elementary proof of the theorem

$$\lim_{x \to \infty} \frac{\vartheta_{k,l}(x)}{x} = \frac{1}{\varphi(k)}, \tag{1.1}$$

where $\varphi(k)$ denotes Euler's function, and

$$\vartheta_{k,l}(x) = \sum_{\substack{p \leq x \\ p \equiv l(k)}} \log p, \tag{1.2}$$

where p denotes the primes, and k and l are integers with $(k,l) = 1$, k positive.

The proof proceeds essentially along the same lines as in a previous paper[1] about the case $k = 1$. However we also need in this case some of the ideas from my paper[2] on Dirichlet's theorem in order to prove that

$$\lim_{x \to \infty} \frac{\vartheta_{k,l}(x)}{x} > 0. \tag{1.3}$$

a result which we will need for our proof of (1.1).

It is possible to shorten the proof in several ways, which however would make it less elementary. For instance one could consider also the complex characters mod k, and in this way avoid Lemma 2 and most of the proof of Lemma 3. Also, by using results about the decomposition of primes in the quadratic field $K(\sqrt{D})$, we could make the proof of Lemma 1 much shorter.

As we shall see, the following proof is completely constructive, in the sense that it gives for any fixed positive ϵ, a way of finding, in a finite number of steps, an x_0—depending on ϵ and k—such that

$$\left| \frac{\vartheta_{k,l}(x)}{x} - \frac{1}{\varphi(k)} \right| < \epsilon.$$

for $x > x_0$.

Actually, it is possible in this way to prove more than (1.1). By careful estimation it is possible to show by the method below that

Received December 30, 1948.

[1] "An Elementary Proof of the Prime-number Theorem," *Ann. of Math.*, vol. 50 (1949), 305-313.

[2] "An Elementary Proof of Dirichlet's Theorem about Primes in an Arithmetic Progression," *Ann. of Math.*, vol. 50 (1949), 297-304.

$$\vartheta_{k,l}(x) = \frac{1}{\varphi(k)}\, x + O\left(\frac{x}{(\log x)^c}\right),$$

where c is a positive constant.

Throughout the paper k denotes a fixed positive integer; l denotes an integer with $(l, k) = 1$, while α, β and γ are used to designate numbers from a reduced residue system mod k; p, q and r denote primes. The letter K denotes positive constants, depending on k only; x_0 denotes a constant (not necessarily the same at each occurrence), depending on k only. In the same way x_σ denotes a number depending only on k and the positive number σ. The constants implied by the O's are generally dependent on k and in secs. 4 and 5 also on σ.

From my two previous papers mentioned above I make use of the prime-number theorem in the case $k = 1$, the formula

$$\sum_{p \leq x} \log^2 p + \sum_{pq \leq x} \log p \log q = 2x \log x + O(x), \tag{1.4}$$

and the further formula

$$\sum_{\substack{p \leq x \\ p \equiv l(k)}} \log^2 p + \sum_{\substack{pq \leq x \\ pq \equiv l(k)}} \log p \log q = \frac{1}{\varphi(k)} \left\{ \sum_{p \leq x} \log^2 p + \sum_{pq \leq x} \log p \log q \right\} + O(x). \tag{1.5}$$

Finally I make use of the well-known formula

$$\sum_{p \leq x} \frac{\log p}{p} = \log x + O(1). \tag{1.6}$$

By partial summation it is easily seen that one may also give (1.6) the forms

$$\sum_{n \leq x} \frac{\vartheta(n)}{n^2} = \log x + O(1). \tag{1.7}$$

$$\sum_{n \leq x} \vartheta\left(\frac{x}{n}\right) = x \log x + O(x), \tag{1.8}$$

where $\vartheta(x)$ denotes $\vartheta_{k,l}(x)$ for $k = 1$.

Trivial estimations are often not carried out in detail, but left to the reader.

2. Fundamental formulas and inequalities. From (1.4) and (1.5) we get

$$\sum_{\substack{p \leq x \\ p \equiv l(k)}} \log^2 p + \sum_{\substack{pq \leq x \\ pq \equiv l(k)}} \log p \log q = \frac{2}{\varphi(k)}\, x \log x + O(x), \tag{2.1}$$

which may also be written[3]

$$\vartheta_l(x) \log x + \sum_{p \leq x} \log p\ \vartheta_{\bar{l}}\left(\frac{x}{p}\right) = \frac{2}{\varphi(k)}\, x \log x + O(x), \tag{2.2}$$

where $\bar{p}$ denotes a solution of the congruence $p\,\bar{p} \equiv 1 \pmod{k}$.

[3] We write $\vartheta_l(x)$ instead of $\vartheta_{k,l}(x)$ where no misunderstanding can occur.

Since

$$\sum_{\substack{p \leqq x \\ p \equiv l(k)}} \log p + \sum_{\substack{pq \leqq x \\ pq \equiv l(k)}} \frac{\log p \log q}{\log pq} = \frac{2}{\varphi(k)} x + O\left(\frac{x}{\log x}\right), \tag{2.3}$$

which follows by partial summation using (2.1), then

$$\begin{aligned}\sum_{\substack{pq \leqq x \\ pq \equiv l(k)}} \log p \log q &= \sum_{p \leqq x} \log p \sum_{\substack{q \leqq x/p \\ q \equiv l\bar{p}(k)}} \log q = \frac{2}{\varphi(k)} x \sum_{p \leqq x} \frac{\log p}{p} \\ &\quad - \sum_{p \leqq x} \log p \sum_{\substack{qr \leqq x/p \\ qr \equiv l\bar{p}(k)}} \frac{\log q \log r}{\log qr} + O\left(x \sum_{p \leqq x} \frac{\log p}{p\,(1 + \log x/p)}\right) \\ &= \frac{2}{\varphi(k)} x \log x - \sum_{qr \leqq x} \frac{\log q \log r}{\log qr} \vartheta_{l\overline{qr}}\left(\frac{x}{qr}\right) + O(x \log\log x).\end{aligned}$$

Inserting this for the second term on the left-hand side of (2.2), we get

$$\vartheta_l(x) \log x = \sum_{pq \leqq x} \frac{\log p \log q}{\log pq} \vartheta_{l\overline{pq}}\left(\frac{x}{pq}\right) + O(x \log\log x). \tag{2.4}$$

Writing now

$$\vartheta_l(x) = \frac{1}{\varphi(k)} x + R_l(x), \tag{2.5}$$

we get from (2.2)

$$R_l(x) \log x = -\sum_{p \leqq x} \log p\, R_{l\overline{p}}\left(\frac{x}{p}\right) + O(x). \tag{2.6}$$

In the same manner (2.4) yields

$$R_l(x) \log x = \sum_{pq \leqq x} \frac{\log p \log q}{\log pq} R_{l\overline{pq}}\left(\frac{x}{pq}\right) + O\,(x \log\log x), \tag{2.7}$$

since

$$\sum_{pq \leqq x} \frac{\log p \log q}{pq \log pq} = \log x + O\,(\log\log x),$$

which follows by partial summation using

$$\sum_{pq \leqq x} \frac{\log p \log q}{pq} = \tfrac{1}{2} \log^2 x + O\,(\log x),$$

which again follows from (1.6).

Combining (2.6) and (2.7) we get

$$2\,|R_l(x)|\log x \leqslant \sum_{p \leqq x} \log p\,\left|R_{l\overline{p}}\left(\frac{x}{p}\right)\right| + \sum_{pq \leqq x} \frac{\log p \log q}{\log pq} \left|R_{l\overline{pq}}\left(\frac{x}{pq}\right)\right| + O(x \log\log x),$$

or if a runs over a reduced residue system mod k,

$$2\,|R_1(x)|\log x \leqslant \sum_a \Bigg\{ \sum_{\substack{p \leqslant x \\ p \equiv l_a(k)}} \log p \left| R_a\left(\frac{x}{p}\right)\right| + \sum_{\substack{pq \leqslant x \\ pq \equiv l_a(k)}} \frac{\log p \log q}{\log pq} \left| R_a\left(\frac{x}{pq}\right)\right| \Bigg\} + O(x \log\log x).$$

From this we get by partial summation, using (2.3),

$$\begin{aligned} 2|R_1(x)|\log x &\leqslant \sum_a \sum_{n \leqslant x} \Bigg\{ \sum_{\substack{p \leqslant n \\ p \equiv l_a(k)}} \log p + \sum_{\substack{pq \leqslant n \\ pq \equiv l_a(k)}} \frac{\log p \log q}{\log pq} \Bigg\} \\ &\quad \cdot \left\{ \left| R_a\left(\frac{x}{n}\right)\right| - \left| R_a\left(\frac{x}{n+1}\right)\right| \right\} + O(x \log\log x) \\ &= \frac{2}{\varphi(k)} \sum_a \sum_{n \leqslant x} n \left\{ \left| R_a\left(\frac{x}{n}\right)\right| - \left| R_a\left(\frac{x}{n+1}\right)\right| \right\} \\ &\quad + O\Bigg(\sum_a \sum_{n \leqslant x} \frac{n}{1+\log n} \left| R_a\left(\frac{x}{n}\right) - R_a\left(\frac{x}{n+1}\right)\right| \Bigg) \\ &\quad + O\,(x \log\log x) = \frac{2}{\varphi(k)} \sum_a \sum_{n \leqslant x} \left| R_a\left(\frac{x}{n}\right)\right| \\ &\quad + O\Bigg(\sum_{n \leqslant x} \frac{x}{n\,(1+\log n)} \Bigg) + O\Bigg(\sum_{n \leqslant x} \frac{n}{1+\log n} \left(\vartheta\left(\frac{x}{n}\right) - \vartheta\left(\frac{x}{n+1}\right)\right)\Bigg) \\ &\quad + O\,(x \log\log x) = \frac{2}{\varphi(k)} \sum_a \sum_{n \leqslant x} \left| R_a\left(\frac{x}{n}\right)\right| \\ &\quad + O\Bigg(\sum_{n \leqslant x} \frac{1}{1+\log n}\, \vartheta\left(\frac{x}{n}\right)\Bigg) + O\,(x \log\log x) \\ &= \frac{2}{\varphi(k)} \sum_a \sum_{n \leqslant x} \left| R_a\left(\frac{x}{n}\right)\right| + O\,(x \log\log x) \end{aligned}$$

or finally[4]

$$|R_1(x)| \leqslant \frac{1}{\varphi(k)\log x} \sum_a \sum_{n \leqslant x} \left| R_a\left(\frac{x}{n}\right)\right| + O\left(\frac{x \log\log x}{\log x}\right). \tag{2.8}$$

3. A lower bound for $\vartheta_1(x)$. From (2.4) in the form

$$\sum_{\substack{p \leqslant x \\ p \equiv l(k)}} \log^2 p = \sum_{\substack{pqr \leqslant x \\ pqr \equiv l(k)}} \frac{\log p \log q \log r}{\log pq} + O\,(x \log\log x)$$

[4]Instead of (2.8) we might use the somewhat sharper inequality

$$|R_1(x)| \leqslant \frac{2}{\varphi(k)\log^2 x} \sum_a \sum_{n \leqslant x} \log n \left| R_a\left(\frac{x}{n}\right)\right| + O\left(\frac{x}{\log x}\right),$$

which can be proved in a similar way.

we get by partial summation

$$\sum_{\substack{p\le x\\ p\equiv l(k)}}\log^2 p\,\log^2\frac{x}{p}=\sum_{\substack{pqr\le x\\ pqr\equiv l(k)}}\frac{\log p\log q\log r}{\log pq}\log^2\frac{x}{pqr}+O(x\log\log x)$$

$$\geqslant\frac{1}{\log x}\sum_{\substack{pqr\le x\\ pqr\equiv l(k)}}\log p\log q\log r\,\log^2\frac{x}{pqr}+O(x\log\log x).$$

Now it is easily seen[5] that if $pqr\leqslant x$,

$$\log^2\frac{x}{pqr}=2\sum_{\substack{p\le\mu,\,q\le\nu\\ \mu\nu\le x/r}}\frac{1}{\mu\nu}+O\left(\left(\frac{1}{p}+\frac{1}{q}\right)\left(1+\log\frac{x}{pqr}\right)\right).$$

Inserting this expression in the preceding inequality, we get

$$\sum_{\substack{p\le x\\ p\equiv l(k)}}\log^2 p\,\log^2\frac{x}{p}$$

$$\geqslant\frac{2}{\log x}\sum_{\substack{pqr\le x\\ pqr\equiv l(k)}}\log p\log q\log r\sum_{\substack{p\le\mu,\,q\le\nu\\ \mu\nu\le x/r}}\frac{1}{\mu\nu}+O(x\log\log x).$$

$$=\frac{2}{\log x}\sum_{\alpha\beta\gamma\equiv l(k)}\sum_{\mu\nu\le x}\frac{1}{\mu\nu}\vartheta_\alpha(\mu)\,\vartheta_\beta(\nu)\,\vartheta_\gamma\left(\frac{x}{\mu\nu}\right)+O(x\log\log x),$$

or

$$\sum_{\substack{p\le x\\ p\equiv l(k)}}\log^2 p\,\log^2\frac{x}{p}\geqslant\frac{2}{\log x}\sum_{\alpha\beta\gamma\equiv l(k)}\sum_{x^{\frac{1}{3}}\le\mu,\nu\le x^{\frac{1}{2}}}\frac{1}{\mu\nu}\vartheta_\alpha(\mu)\vartheta_\beta(\nu)\vartheta_\gamma\left(\frac{x}{\mu\nu}\right)+O(x\log\log x).\tag{3.1}$$

LEMMA 1. *If χ is a real non-principal character mod k, then for $x>x_0$ we have*

$$\sum_{\chi(\alpha)=1}\vartheta_\alpha(x)>K_1x,\qquad\sum_{\chi(\alpha)=-1}\vartheta_\alpha(x)>K_1x.$$

It is obviously sufficient if we prove that

$$\sum_{\substack{p\le x\\ \chi(p)=1}}\frac{\log p}{p}=\tfrac{1}{2}\log x+O(1),\tag{3.2}$$

because then, if $0<\delta<1$ is a fixed number, we get

$$\sum_{\chi(\alpha)=1}\vartheta_\alpha(x)>\delta x\sum_{\substack{\delta x\le p\le x\\ \chi(p)=1}}\frac{\log p}{p}=\tfrac{1}{2}\left(\delta\log\frac{1}{\delta}\right)x+O(\delta x)>K_1x,$$

if δ is chosen small enough and $x>x_0$. The second part of the lemma follows in the same way by combining (1.6) and (3.2).

---

[5] For example, by noting that $\sum_{q\le\nu\le x/\mu r}\frac{1}{\nu}=\sum_{\mu\le\nu\le x/qr}\frac{1}{\nu}+O\left(\frac{1}{q}+\frac{1}{\mu}\right)$.

To prove (3.2) we remark that[6] to each such character χ there corresponds an integer D, which is not a square, with $|D| < k^2$, and such that $\chi(p) = (D|p)$ for all primes p. Here $(D|p)$ is the usual quadratic residue symbol. Hence (3.2) is equivalent to

$$\sum_{\substack{p \leqq x \\ (D|p)=1}} \frac{\log p}{p} = \tfrac{1}{2}\log x + O(1). \tag{3.3}$$

To prove (3.3) we consider the product

$$P = \prod_{\substack{|u| \leqslant \sqrt{x/2} \\ |v| \leqslant \sqrt{x/2|D|}}}{}' |u^2 - Dv^2|, \tag{3.4}$$

where the dash Π' indicates that the term $u = v = 0$ is omitted. It is easily seen that[7]

$$\log P = \frac{2x}{\sqrt{|D|}} \log x + O(x). \tag{3.5}$$

Let us estimate the highest power dividing P of a prime $p \leqslant x$.

First assume that $(D|p) = 1$. We first estimate how many solutions the congruence

$$u^2 - Dv^2 \equiv 0 \pmod{p}, \tag{3.6}$$

has in the given range for u and v. Since $(D|p) = 1$ there clearly exist solutions of (3.6) which are nontrivial i.e. with $(u, p) = (v, p) = 1$. Let now u_0, v_0 be a fixed such nontrivial solution. Then if u, v also is a solution we have

$$(uv_0)^2 - (u_0v)^2 \equiv 0 \pmod{p},$$

or one of the congruences

$$u v_0 \mp u_0 v \equiv 0 \pmod{p}, \tag{3.7}$$

must be satisfied. Conversely a solution u, v of (3.7) is a solution of (3.6). Consider (3.7) with the upper sign. Obviously the "vectors" (u, v) satisfying (3.7) form a two-dimensional lattice. Since there exist integers (u, v) with $uv_0 - u_0v = p$, the area of a "period-parallelogram" or single "cell" in the lattice is p, because it obviously could not be less than p. Also no "vector" in the lattice has a length less than $\sqrt{p/|D|}$, since for $(u, v) \neq (0,0)$ we have

$$u^2 + v^2 \geqslant |u^2 - Dv^2|/|D| \geqslant p/|D|.$$

From this it is easily seen[8] that the lattice has a basis[9] of two vectors both $< 2\sqrt{|D|p}$, and hence that the rectangle $|u| \leqslant \sqrt{x/2}$, $|v| \leqslant \sqrt{x/2|D|}$ with area $2x/\sqrt{|D|}$ contains

---

[6] See for instance Dirichlet-Dedekind: *Vorlesungen über Zahlentheorie* (the beginning of §135).

[7] For example, by showing that the number of terms with $|u^2 - Dv^2| \leqslant T$ is $O(\sqrt{xT})$.

[8] For example, by noting that a "period-parallelogram" may always be chosen so that neither of its sides is greater than a diagonal.

[9] Or, otherwise expressed, that the lattice may be built up of "period-parallelograms" with both sides $< 2\sqrt{|D|p}$.

$$\frac{2x}{p\sqrt{|D|}}+O\left(\sqrt{\frac{x}{p}}\right)$$

such lattice points (u, v). Hence we have as many solutions of (3.7) with the upper sign.

Treating the case of the lower sign in the same way, we get altogether

$$\frac{4x}{p\sqrt{|D|}}+O\left(\sqrt{\frac{x}{p}}\right)+O\left(\frac{x}{p^2}\right),$$

solutions of (3.6) in the given range for u and v, because the two congruences (3.7) only have common solutions with $u \equiv v \equiv 0 \pmod{p}$, the number of which is $O(x/p^2)$ since we exclude the solution $u = v = 0$.
Thus

$$\frac{4x}{p\sqrt{|D|}}+O\left(\sqrt{\frac{x}{p}}\right)+O\left(\frac{x}{p^2}\right)$$

of the factors $|u^2 - Dv^2|$ of P contain p as a factor.

In the same way we find that $O(x/p^i)$ of them contain p^i as a factor for $i>1$. Thus the highest power of p dividing P has the exponent

$$\frac{4x}{p\sqrt{D}}+O\left(\sqrt{\frac{x}{p}}\right)+O\left(\frac{x}{p^2}\right).$$

On the other hand, if $(D|p) = -1$, we see that p has to divide both u and v in order to divide $|u^2 - Dv^2|$. From this it is easily seen that P in this case contains p only to a power with exponent $O(x/p^2)$.

Finally if $(D|p) = 0$ or p divides D, we see that P contains p to a power with exponent $O(x/p)$. Combining these results we get

$$\log P = \frac{4x}{\sqrt{|D|}}\sum_{\substack{p\\(D|p)=1}}^{\leqq x}\frac{\log p}{p}+O\left(\sqrt{x}\sum_{p\leqq x}\frac{\log p}{\sqrt{p}}\right)+O\left(x\sum_{p\leqq x}\frac{\log p}{p^2}\right)$$
$$+O\left(x\sum_{p/D}\frac{\log p}{p}\right)=\frac{4x}{\sqrt{|D|}}\sum_{\substack{p\\(D|p)=1}}^{\leqq x}\frac{\log p}{p}+O(x).$$

Comparing this with (3.5) we get (3.3), which proves our lemma.

LEMMA 2. *Let $h = \frac{1}{2}\varphi(k)$, and suppose that we have three sets of h different residues[10] mod k: $\alpha_1, \alpha_2, \ldots, \alpha_h$; $\beta_1, \beta_2, \ldots, \beta_h$; $\gamma_1, \gamma_2, \ldots, \gamma_h$. Further suppose that for each non-principal real character χ mod k there is at least one α_i with $\chi(\alpha_i) = 1$, and at least one with $\chi(\alpha_i) = -1$. Then there exists a triple α, β, γ belonging to the sets, such that $\alpha\beta\gamma \equiv l \pmod{k}$.*

Suppose that always $\alpha\beta\gamma \not\equiv l \pmod{k}$, or $\alpha\beta \not\equiv l\overline{\gamma} \pmod{k}$. This implies that there are h different values[11] the product $\alpha\beta$ cannot assume, or since $h = \frac{1}{2}\varphi(k)$, that the product $\alpha\beta$ can assume only h different values. Writing $\alpha_i = \alpha_1\alpha'_i$ and $\beta_i = \beta_1\beta'_i$ for $i = 1, 2, 3, \ldots, h$, this means that the products $\alpha'\beta'$ can assume only h different values. Since $\alpha'\beta'$ can assume the values 1,

[10] By residues, we understand here residues belonging to the reduced residue system.
[11] By values we mean here residues mod k.

$\alpha'_2, \ldots, \alpha'_h$ and also the values $1, \beta'_2, \ldots, \beta'_h$, the sets α' and β' are identical, and it follows easily that the set $1, \alpha'_2, \ldots, \alpha'_h$, forms a group with respect to multiplication. Now we define a character χ having the value 1 for all residues of the set α', and the value -1 for all remaining residues of the reduced residue system. Then we have $\chi(a_i) = \chi(a_1)\chi(\alpha'_i) = \chi(a_1)$, which contradicts the assumption that the set a contains both a_i's with $\chi(a_i) = 1$, and such with $\chi(a_i) = -1$. This proves our lemma.

LEMMA 3. *We have for* $x > x_0$,

$$\vartheta_1(x) > K_3 x.$$

From (2.3) follows for all a and $x > x_0$,

$$\vartheta_a(x) \leqslant \frac{2}{\varphi(k)} x + O\left(\frac{x}{\log x}\right) < \frac{2}{\varphi(k)}\left(1 + \frac{1}{2\varphi(k)}\right) x. \tag{3.8}$$

Also from the prime-number theorem in the case $k = 1$, we get for $x > x_0$,

$$\sum_a \vartheta_a(x) = x + o(x) > \left(1 - \frac{1}{2\varphi(k)}\right) x. \tag{3.9}$$

The inequality

$$\vartheta_a(x) > \frac{x}{\varphi^2(k)}$$

must hold for at least $h = \frac{1}{2}\varphi(k)$ values of a. For if $\vartheta_a(x) \leqq x/\varphi^2(k)$ for $m > h$ values of a then, by (3.8) and (3.9),

$$\left(1 - \frac{1}{2\varphi(k)}\right) x < \sum_a \vartheta_a(x) < \left(\frac{m}{\varphi^2(k)} + \frac{2(\varphi(k) - m)}{\varphi(k)}\left(1 + \frac{1}{2\varphi(k)}\right)\right) x,$$

which leads to a contradiction.

Also, from Lemma 1, there is at least one a with $\chi(a) = 1$ and at least one with $\chi(a) = -1$ satisfying

$$\vartheta_a(x) > \frac{2K_1 x}{\varphi(k)}.$$

Hence there exists a set of $h = \frac{1}{2}\varphi(k)$ different residues $a_1, a_2, \ldots, a_h \bmod k$, for each μ in the range $x^{\frac{1}{4}} \leqslant \mu \leqslant x^{\frac{1}{4}}$, $x > x_0$ such that

$$\vartheta_{a_i}(\mu) > K_3 \mu, \tag{3.10}$$

for $i = 1, 2, \ldots, h$ where

$$K_3 = \min\left(\frac{2}{\varphi(k)} K_1, \frac{1}{\varphi^2(k)}\right),$$

and such that for any real non-principal character $\chi \bmod k$ there is an a_i in the set with $\chi(a_i) = 1$, and another with $\chi(a_i) = -1$.

Arguing in the same way for $\vartheta_\beta(\nu)$ and $\vartheta_\gamma(x/\mu\nu)$ where $x^{\frac{1}{4}} \leqslant \mu \leqslant x^{\frac{1}{4}}$, $x^{\frac{1}{4}} \leqslant \nu \leqslant x^{\frac{1}{4}}$, $x > x_0$, we find from Lemma 2 that for each pair μ, ν in the ranges $x^{\frac{1}{4}} \leqslant \mu \leqslant x^{\frac{1}{4}}$, $x^{\frac{1}{4}} \leqslant \nu \leqslant x^{\frac{1}{4}}$ for $x > x_0$ there is a triple α, β, γ with

$$\vartheta_\alpha(\mu) > K_2\mu,\ \vartheta_\beta(\nu) > K_2\nu,\ \vartheta_\gamma(x/\mu\nu) > K_2 x/\mu\nu,$$

and $\alpha\beta\gamma \equiv l \pmod{k}$. Hence (3.1) gives

$$\sum_{\substack{p \leqq x \\ p \equiv l(k)}} \log^2 p \log^2 \frac{x}{p} > \frac{2}{\log x} K_2^3 x \left(\sum_{x^{\frac{1}{3}} < \mu \leqq x^{\frac{1}{2}}} \frac{1}{\mu} \right)^2 + O(x \log\log x)$$

$$= \frac{2}{\log x} K_2^3 x \left(\tfrac{1}{12} \log x\right)^2 + O(x \log\log x) > K_4 x \log x,$$

for $x > x_0$.

Thus if $\delta < 1$ is a fixed positive number, we get

$$\log x \log^2 1/\delta \,.\, \vartheta_l(x) \geqslant \sum_{\substack{p \leqq x \\ p \equiv l(k)}} \log^2 p \log^2 x/p$$

$$- \sum_{p \leqq \delta x} \log^2 p \log^2 x/p > K_4 x \log x$$

$$- K_5 \delta x \log x \log^2 1/\delta = (K_4 - K_5 \delta \log^2 1/\delta)\, x \log x,$$

or if δ is chosen small enough,

$$\vartheta_l(x) > K_3 x,$$

for $x > x_0$, which proves Lemma 3.

From (2.2) we get, using Lemma 3,

$$\vartheta_l(x) \leqslant \frac{2}{\varphi(k)} x - \frac{K_3 x}{\log x} \sum_{p \leqq x} \frac{\log p}{p} + O\left(\frac{x}{\log x}\right),$$

or

$$\vartheta_l(x) < \left(\frac{2}{\varphi(k)} - K_6\right) x,$$

for $x > x_0$. This combined with Lemma 3 and (2.5) gives

$$|R_l(x)| < \frac{\sigma_0}{\varphi(k)} x, \tag{3.11}$$

for $x > x_0$, where $\sigma_0 < 1$ is a positive number depending on k only.

4. Properties of $R_l(x)$. From (2.1) we get by partial summation

$$\sum_{\substack{p \leqq x \\ p \equiv l(k)}} \log^2 p \log \frac{x}{p} + \sum_{\substack{pq \leqq x \\ pq \equiv l(k)}} \log p \log q \log \frac{x}{pq} \tag{4.1}$$

$$= \frac{2}{\varphi(k)} \sum_{n \leqq x} \log n \log \frac{x}{n} + O(x) = \frac{2}{\varphi(k)} x \log x + O(x).$$

Now

$$\log \frac{x}{p} = \sum_{p < n \leqq x} \frac{1}{n} + O\left(\frac{1}{p}\right),$$

and

$$\log \frac{x}{pq} = \sum_{p < n \leqq x/q} \frac{1}{n} + O\left(\frac{1}{p}\right).$$

Inserting this in (4.1), we get

$$\sum_{n \leqslant x} 1/n \sum_{\substack{p \leqslant n \\ p \equiv l(k)}} \log^2 p + \sum_{n \leqslant x} 1/n \sum_{p \leqslant n} \log p \sum_{\substack{q \leqslant x/n \\ q \equiv l'p(k)}} \log q$$

$$= \frac{2}{\varphi(k)} x \log x + O(x),$$

which may be written in the form

$$\sum_{n \leqslant x} \frac{\log n}{n} \vartheta_l(n) + \sum_{n \leqslant x} \frac{1}{n} \sum_{\alpha\beta \equiv l(k)} \vartheta_\alpha(n) \vartheta_\beta\left(\frac{x}{n}\right) = \frac{2}{\varphi(k)} x \log x + O(x).$$

This gives

$$\sum_{n \leqslant x} \frac{\log n}{n} R_l(n) + \frac{1}{\varphi(k)} x \log x + O(x) = \frac{1}{\varphi(k)} x \log x$$

$$- \sum_{n \leqslant x} \frac{1}{n} \sum_{\alpha\beta \equiv l(k)} R_\alpha(n) R_\beta\left(\frac{x}{n}\right) - \frac{1}{\varphi(k)} \sum_{n \leqslant x} \sum_{\alpha} R_\alpha\left(\frac{x}{n}\right)$$

$$- \frac{x}{\varphi(k)} \sum_{n \leqslant x} \frac{1}{n^2} \sum_{\alpha} R_\alpha(n) + O(x),$$

or, using (1.7) and (1.8), and noticing that

$$\sum_{\alpha} R_\alpha(y) = \vartheta(y) - y + O(1),$$

we get finally

$$\sum_{n \leqslant x} \frac{\log n}{n} R_l(n) = - \sum_{n \leqslant x} \frac{1}{n} \sum_{\alpha\beta \equiv l(k)} R_\alpha(n) R_\beta\left(\frac{x}{n}\right) + O(x). \tag{4.2}$$

Suppose now that for a positive fixed number $\sigma \leqslant \sigma_0$, we have for $x > x_\sigma$,

$$|R_l(x)| < \frac{\sigma}{\varphi(k)} x, \tag{4.3}$$

for all $(l, k) = 1$. (4.2) then yields

$$\left| \sum_{n \leqslant x} \frac{\log n}{n} R_l(n) \right| < \frac{\sigma^2}{\varphi(k)} x \sum_{n \leqslant x} \frac{1}{n} + O(x) = \frac{\sigma^2}{\varphi(k)} x \log x + O(x)$$

or if $x_1 < x$,

$$\left| \sum_{x_1 \leqslant n \leqslant x} \frac{\log n}{n} R_l(n) \right| < \frac{\sigma^2}{\varphi(k)} (x + x_1) \log x + O(x).$$

Now let

$$x_1 = (1 - \sigma)^{16} x < \frac{1 - \sigma}{1 + 15\sigma} x.$$

If $R_l(n)$ does not change sign in the interval $x_1 < n \leqslant x$, then the above inequality implies that there exists a y in the interval $x_1 < y \leqslant x$, such that

$$\left|\frac{R_l(y)}{y}\right| \sum_{x_1 \leqslant n \leqslant x} \log n < \frac{\sigma^2}{\varphi(k)} (x + x_1) \log x + O(x),$$

or

$$\left|\frac{R_l(y)}{y}\right| (x - x_1) \log x < \frac{\sigma^2}{\varphi(k)} (x + x_1) \log x + O(x),$$

and

$$\left|\frac{R_l(y)}{y}\right| < \frac{\sigma^2}{\varphi(k)} \frac{x + x_1}{x - x_1} + O\left(\frac{1}{\log x}\right) \tag{4.4}$$

$$< \frac{1}{\varphi(k)} \frac{\sigma(1 + 7\sigma)}{8} + O\left(\frac{1}{\log x}\right) < \frac{1}{\varphi(k)} \frac{\sigma(1 + 3\sigma)}{4},$$

for $x > x_\sigma$. Obviously there exists such a y also in the case that $R_l(n)$ changes sign[13] in the interval $x_1 < n \leqslant x$.

For $y_1 < y_2$, it follows from (2.3) that

$$0 \leqslant \sum_{\substack{y_1 \leqslant p \leqslant y_2 \\ p \equiv l(k)}} \log p \leqslant \frac{2}{\varphi(k)} (y_2 - y_1) + O\left(\frac{y_2}{\log y_2}\right),$$

or

$$|R_l(y_2) - R_l(y_1)| < \frac{1}{\varphi(k)} (y_2 - y_1) + O\left(\frac{y_2}{\log y_2}\right),$$

so that if $\frac{1}{2} \leqslant y'/y \leqslant 2$, and $x_1 < y \leqslant x$, $x_1 < y' \leqslant x$, we get

$$|R_l(y') - R_l(y)| \leqslant \frac{1}{\varphi(k)} |y' - y| + O\left(\frac{y}{\log x}\right).$$

Thus

$$|R_l(y')| < |R_l(y)| + \frac{1}{\varphi(k)} (y' - y) + O\left(\frac{y}{\log x}\right),$$

or

$$\left|\frac{R_l(y')}{y'}\right| < \left|\frac{R_l(y)}{y}\right| \frac{y}{y'} + \frac{1}{\varphi(k)} \left|1 - \frac{y}{y'}\right| + O\left(\frac{1}{\log x}\right).$$

Now let

$$e^{-\delta} \leqslant \frac{y'}{y} \leqslant e^{\delta}$$

where

$$\delta = \frac{\sigma(1 - \sigma)}{32}.$$

The above inequality then gives, by (4.4),

[13] For there is then a y in the interval with $|R_l(y)| < \log y$.

$$\left|\frac{R_l(y')}{y'}\right| < \frac{1}{\varphi(k)}\,\frac{\sigma(1+3\sigma)}{4}\,e^{\delta} + \frac{1}{\varphi(k)}\,(e^{\delta}-1) + O\left(\frac{1}{\log x}\right)$$
$$< \frac{1}{\varphi(k)}\,\frac{\sigma(3+5\sigma)}{8} + O\left(\frac{1}{\log x}\right) < \frac{1}{\varphi(k)}\,\frac{\sigma(1+\sigma)}{2},$$

for $x > x_\sigma$. Thus we have proved, assuming (4.3),

LEMMA 4. *For $x > x_\sigma$, any interval $((1-\sigma)^{16}x, x)$ contains a sub-interval $(y, e^{\delta}y)$ where $\delta = \dfrac{\sigma(1-\sigma)}{32}$, such that for all $y \leqslant z \leqslant e^{\delta}y$ we have*

$$\left|\frac{R_l(z)}{z}\right| < \frac{1}{\varphi(k)}\,\frac{\sigma+\sigma^2}{2}.$$

5. Proof of the prime-number theorem for the arithmetic progression. We shall now prove the

THEOREM.

$$\lim_{x\to\infty}\frac{\vartheta_l(x)}{x} = \frac{1}{\varphi(k)}.$$

Obviously this is equivalent to

$$\lim_{x\to\infty}\frac{R_l(x)}{x} = 0. \tag{5.1}$$

We have that for all $x > 1$ and $(k,l) = 1$,

$$|R_l(x)| < K_7 x, \tag{5.2}$$

and from (3.11), that for $x > x_0$,

$$|R_l(x)| < \frac{\sigma_0}{\varphi(k)}\,x, \tag{5.3}$$

where $\sigma_0 < 1$.

Now assume as in the preceding paragraph, that for a fixed positive number $\sigma \leqslant \sigma_0$, we have for all l,

$$|R_l(x)| < \frac{\sigma}{\varphi(k)}\,x, \tag{5.4}$$

for $x > x_\sigma$. Writing further $\rho = (1-\sigma)^{-16}$, we have then from Lemma 4, that for $x > x_\sigma$ and all a each interval (ρ^{r-1}, ρ^r) where $\rho \leqslant \rho^r \leqslant x/x_\sigma$ contains a sub-interval (y_r, y'_r) with $y'_r = e^{\delta}y_r$ and $\delta = \dfrac{\sigma(1-\sigma)}{32}$, such that for $y_r \leqslant n \leqslant y'_r$ we have

$$\left|\frac{n}{x}\,R_a\left(\frac{x}{n}\right)\right| < \frac{1}{\varphi(k)}\,\frac{\sigma+\sigma^2}{2}.$$

(2.8) then yields

$$
\begin{aligned}
|R_l(x)| &< \frac{1}{\varphi(k)\log x}\sum_{a}\sum_{n\leqslant z}|R_a\left(\frac{x}{n}\right)| + O\left(\frac{x}{\sqrt{\log x}}\right)\\
&< \frac{K_7 x}{\log x}\sum_{z/z_\sigma\leqslant n\leqslant z}\frac{1}{n} + \frac{\sigma x}{\varphi(k)\log x}\sum_{n\leqslant z/z_\sigma}\frac{1}{n}\\
&\quad - \frac{(\sigma-\sigma^2)x}{2\varphi^2(k)\log x}\sum_{a}\sum_{\rho'\leqslant z/z_\sigma}\sum_{y_\rho\leqslant n\leqslant y'_\rho}\frac{1}{n} + O\left(\frac{x}{\sqrt{\log x}}\right)\\
&= \frac{\sigma}{\varphi(k)}x - \frac{(\sigma-\sigma^2)x}{2\varphi(k)\log x}\sum_{\rho'\leqslant z/z_\sigma}\delta + O\left(\frac{x}{\sqrt{\log x}}\right)\\
&= \frac{\sigma}{\varphi(k)}x - \frac{\delta\sigma(1-\sigma)}{2\varphi(k)\log\rho}x + O\left(\frac{x}{\sqrt{\log x}}\right)\\
&= \frac{\sigma}{\varphi(k)}\left(1-\frac{\sigma(1-\sigma)^2}{1024\log 1/1-\sigma}\right)x + O\left(\frac{x}{\sqrt{\log x}}\right)\\
&< \frac{\sigma}{\varphi(k)}\left(1-\frac{(1-\sigma)^2}{2000}\right)x,
\end{aligned}
$$

for $x > x_\sigma$.

Since the iteration-process

$$\sigma_{n+1} = \sigma_n\left(1-\frac{(1-\sigma_n)^2}{2000}\right),$$

starting with $0 < \sigma_0 < 1$, converges to zero (one sees easily that $\sigma_n < e^{-K_8 n}$), this proves (5.1) and hence our theorem.

The Institute for Advanced Study
and
Syracuse University

STAR DIAGRAMS AND THE SYMMETRIC GROUP

R. A. STAAL

Introduction. The irreducible representations of the symmetric group S_n, were shown by A. Young to be in one-to-one correspondence with certain arrays of n nodes. E.g. for $n = 12$ and the partition $\lambda = [4, 4, 3, 1]$ we have the array

$$\lambda: \begin{array}{l} \cdot\,\cdot\,\cdot\,\cdot \\ \cdot\,\cdot\,\cdot\,\cdot \\ \cdot\,\cdot\,\cdot \\ \cdot \end{array}$$

which we call a "Young diagram." The question arises as to the manner in which various properties of the representations are reflected in their corresponding Young diagrams.

The study of modular representations [1] has shown that, relative to a given prime, p, the ordinary (non-modular) irreducible representations of a group gather into "p-blocks". Two irreducible representations of S_n belong to the same p-block if and only if their corresponding diagrams have the same "p-core" (see 1.8). This was conjectured by T. Nakayama in 1940 [3], and proven by R. Brauer and G. de B. Robinson in 1947 [2]. The proof involved an auxiliary diagram—the "star diagram" of the Young diagram concerned. It is the purpose of the present paper to discuss the construction of the star diagram in greater detail, and to place greater emphasis upon it than has hitherto been done.

The distribution into p-blocks has to do with the power, $e(z)$, to which p divides the degree, z, of the representation concerned. Nakayama [4] obtained a formula for $e(z)$ in terms of the diagram's "p-series" (see 1.7), namely

$$e(z) = e(n!) - \sum_i e(p^{a_i}!).$$

This formula, however, was not suitable for the proof of his conjecture. Nakayama was led to the p-series of λ by the study of the "p-hook" (see 1.4) structure of λ. Robinson [5] showed (see below) that this p-hook structure could be represented by an associated diagram, λ^*—the "star diagram" of λ (usually denoted λ_p^*: we omit the subscript, reserving the space for another use later on). The diagram λ^* is in general skew (see 1.2) and to such a diagram corresponds a reducible representation [6] of S_m, where m is the number of nodes of the diagram. The following formula [2] was used in the proof of the conjecture:

Received January 26, 1949. (This paper contains the substance of a thesis prepared under the supervision of Professor G. de B. Robinson and submitted for a Ph.D. degree at the University of Toronto in October, 1948.)

A: $$e(z_\lambda) = e(n!) - e((n-A)!) + e(z_{\lambda^*}).$$

(A denotes the number of nodes of the p-core of λ, and z_{λ^*} is the degree of the reducible representation corresponding to λ^*.)

The proof of formula **A**, however, was based on Nakayama's formula, which involved the p-series, —an entity not appearing in A. Accordingly it was felt that full use was not being made of the star diagram, λ^*, and it was hoped that a proof could be developed in terms of it alone.

The present paper begins with a proof of the following existence theorem for λ^*:

B: *Given a right diagram, λ, and a positive integer, q, there exists a diagram, λ^*, such that there is a one-to-one correspondence between kq-hooks of λ and k-hooks of λ^*.*

An auxiliary theorem, **B′**, shows that λ^* represents the actual q-hook *structure* of λ, with regard to removal of q-hooks from λ. Simple considerations of congruence provide a new proof of the fact that λ^* has at most q disjoint constituents.

The following theorem exhibits the connection between λ and λ^* in a form which leads to a new proof of **A**:

C: *Gather the δ's of λ (see 1.9) into classes of δ's which are congruent (mod q). For each such class form the diagram whose δ's are those of this class. The star diagrams of the diagrams thus formed are the constituents of λ^*.*

A proof of **A** is then given which depends only on λ^*, and pairs the factors p of z_λ and z_{λ^*} in an explicit manner. (In **A** we take q to be a prime, p.)

The diagram λ^* and the p-series of λ are shown to be related in the following way:

D: *Given λ, form λ^*, $(\lambda^*)^* = \lambda^{2*}, \ldots, \lambda^{r*}$, where λ^{r*} is a p-core. Suppose the p-core of λ^{i*} has A_i nodes. Then the p-series of λ is*

$$p^r, \ldots (A_r \text{ times}); p^{r-1}, \ldots (A_{r-1} \text{ times}); \ldots \quad ; p, \ldots (A_1 \text{ times}).$$

This leads to a new proof of Nakayama's p-series formula for $e(z)$, based on Theorem **A**. All told, emphasis on the star diagram, rather than the p-series, is seen to recast the theory in a more orderly and understandable manner.

1. Notation, definitions. We shall find it convenient to collect the numerous definitions which are required into a preliminary section.

1.1 A *right* (Young) *diagram* is an array of n nodes with straight top and left sides, and whose rows are in non-increasing order of length. A node which has no node one row below and one column to the right of it is said to be a node of the *rim* of the diagram.

1.2 A ***skew diagram*** is what is obtained by removing from the top left corner of a right diagram another right diagram which is contained in it. E.g.

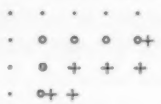
1.21

The circled nodes form a skew diagram.

1.3 A *disjoint* (*skew*) *diagram* is one which consists of constituents having no rows or columns with nodes in common. (See star diagram of 2.1).

1.4 A ***right hook*** of a diagram, λ, consists of a node of λ, together with all nodes directly below it, and directly to the right of it. If it has q nodes, it is called a *q-hook*, or a hook of *length q*. We shall call the top-right and bottom-left nodes of a hook its *top* and *bottom* nodes, respectively. E.g.

1.41



The circle nodes form a 6-hook. (Two of them are marked "+" as well).

1.5 Each right hook has an associated ***skew hook***, of the same length, consisting of all the nodes along the rim from the top node of the given hook to its bottom node. (The nodes in 1.41 marked "+" form a skew hook.) A piece of the rim is the skew hook of some right hook if, and only if, its top node has no node to the right of it (in its row) and its bottom node has no node below it (in its column).

1.6 To ***remove a hook*** from λ, we erase the nodes of its associated skew hook. (The removal of the 6-hook of 1.41 leaves the diagram whose nodes are those not marked "+".)

1.7 The *p-series* of λ: Let the longest hook of λ whose length is a power of p be of length p^{e_1}. Remove this hook from λ. Let the longest hook of the remaining diagram whose length is a power of p, but not greater than p^{e_1}, be of length p^{e_2}. Remove this hook and repeat the process until all such hooks are removed. The resulting sequence,

$$p^{e_1}, p^{e_2}, p^{e_3}, \ldots$$

is the p-series of λ. (The 2-series of 1.81 is 8,4.)

1.8 The *p-core* of λ: The result of the successive removal of the hooks of the p-series is a right diagram—the p-core of λ. It is uniquely determined by p

and λ, and is obtained also when all kp-hooks ($k = 1,2,3,\ldots$) are removed from λ in any order whatever [4]. E.g.

1.81

The 2-core consists of a single node (circle).

1.9 The *δ-numbers* of λ: The lengths of the hooks which begin in the top row of λ are called the "δ-numbers" or "δ's" of λ. They are numbered in the order of their lengths, i.e. $\delta_1 > \delta_2 > \delta_3 > \ldots$. "$\delta$" is also used to refer to the hook whose length is δ. This convenient ambiguity causes no difficulties. (The δ's of 1.21 are 9, 7, 6, 4, 1 and those of 1.41 are 8, 7, 6, 4, 3.)

2. Star diagrams.

We prove the following existence theorem.

B. *Given the right diagram λ, and a positive integer q, there exists a diagram λ^* (called the "star diagram" of λ) such that there is a one-to-one correspondence between kq-hooks of λ and k-hooks of λ^*.*

Example:

2.1 $q = 2$ λ:  λ^*: 

It is sufficient and simpler to consider skew hooks (see 1.5) and we shall now refer to these simply as "hooks". We shall consider the nodes of a hook to be ordered from top to bottom and right to left. A node which has no node to the right of it (in its row) we shall call an "H", and a node which has no node below it (in its column)we shall call an "F". An F cannot precede (immediately) an H. If a node is not an F, then the node which follows it is an H.

Carrying out Robinson's construction[5], we take the longest (any one of the longest) kq-hook, J_1, in λ, of length k_1q, say, and consider the chain, C_1, of all kq-hooks having the same top node, H_1, as J_1. Suppose the lengths of these kq-hooks are k_1q, k_2q, k_3q, $\ldots$ k_sq; we construct the diagram whose δ's are $k_1, k_2, \ldots, k_s$ ($k_1 > k_2 > k_3 > \ldots > k_s$.) This diagram we take to be the first constituent, λ^*_1, of λ^*. (In 2.1, $k_1 = 3$, $k_2 = 2$.) This gives a one-to-one correspondence between the kq-hooks of C_1 and k-hooks of λ^*_1 which begin in the top row of λ^*_1. More than this follows, however.

Consider *any* k-hook, K, of λ^*_1, consisting of the $(r+1)$th, $\ldots$, $(r+k)$th nodes of the rim of λ^*_1. Its bottom node is the bottom node of an $(r+k)$-hook of λ^*_1, beginning in the top row of λ^*_1: this hook corresponds to an $(r+k)q$-hook of λ, beginning at H_1, and whose bottom node, an F, is the $(r+k)q$th node of J_1.

Suppose the rqth node of J_1 were to the right of the $(rq+1)$th node. Then it would be an F, and the first rq nodes of J_1 would form an rq-hook corresponding to an r-hook consisting of the first r nodes of the rim of λ^*_1. But the $(r+1)$th node of the rim of λ^*_1 is the top node of the hook K, and such a node cannot follow the bottom node of a hook. Hence the $(rq+1)$th node of J_1 is an H.

Therefore, corresponding to K there is a kq-hook of λ, beginning at the $(rq+1)$th node of J_1.

Next, let I be a kq-hook of λ whose bottom node is the tqth node of J_1, which is the bottom node of a hook of C_1. The tth node of the rim of λ^*_1 must then be an F; k cannot exceed t, for otherwise J_1 would not be the longest hook of λ whose length is a multiple of q. The top node of I is an H—hence the preceding node is not an F, and the first $(t-k)q$ nodes of J_1 do not form a hook. Hence the first $(t-k)$ nodes of the rim of λ^*_1 do not form a hook, and the $(t-k+1)$th node is an H.

Hence the k nodes of the rim of λ^*_1 which follow the $(t-k)$th node form a k-hook corresponding to I.

So far we have a one-to-one correspondence between k-hooks of λ^*_1 *and kq-hooks of* λ *whose bottom nodes are bottom nodes of hooks of* C_1. For each k-hook of λ^*_1 there is a δ-hook of λ^*_1 having the same bottom node: the corresponding kq-hook of λ has the same bottom node as the hook of λ corresponding to this δ-hook.

Continuing Robinson's construction, we take the longest kq-hook, J_2, of λ *which is not already represented* in λ^*_1 and repeat the previous construction, obtaining λ^*_2. We continue in this way until all the kq-hooks of λ are used up, obtaining diagrams $\lambda^*_1, \lambda^*_2, \ldots \lambda^*_w$, which we arrange disjointly to form λ^*.

It remains to show that a given kq-hook, M, of λ is represented in only one constituent of λ^*. Let λ^*_a, λ^*_b, $a \neq b$, correspond to the chains C_a, C_b of kq-hooks. It is sufficient to show that the bottom node of M cannot be the bottom node of a hook of C_a and a hook of C_b.

Suppose it were, and suppose the top node of J_a were m nodes above the top node of J_b. Then m would be divisible by q, and the hook running from the top of J_a to the bottom of J_b would belong to C_a. But then J_b would be represented in λ^*_a, and also $a > b$, since J_a would be longer than J_b: thus J_b would already have been represented, contrary to hypothesis.

Hence the correspondence is one-to-one, and the theorem is proven.

The following theorem shows that if corresponding hooks are removed from λ and λ^*, the relationship between them is unaltered.

B′. *If a k-hook is removed from* λ^*, *leaving* $\bar{\lambda}^*$, *and the corresponding kq-hook is removed from* λ, *leaving* $\bar{\lambda}$, *then* $(\bar{\lambda})^* = \bar{\lambda}^*$.

Nakayama showed that the removal of a kq-hook can be accomplished by the successive removal of k q-hooks. Hence it is sufficient to consider the removal of a single node from λ^*. This will affect only the constituent, λ^*_i,

in which it appears. If it is the top node of λ^*_i, it will (when removed) reduce all the δ's of that constituent by 1: if it is not the top node, it will reduce exactly one δ by 1—namely the δ of which it is the bottom node. Let B be the q-hook of λ corresponding to the node removed from λ^*, and let h,f be its top and bottom nodes respectively. Consider the effect of removing B from λ.

When B is removed, the node (if there is one) preceding h (on the rim of λ) becomes an F. (It was not previously an F.) Any other nodes preceding B remain unaffected. The node following f (if there is one) becomes an H. (It was not previously an H.) All other nodes following B remain unaffected. If there are nodes preceding B and nodes following B, then when B is removed, q new nodes become members of the rim in its place. These are the nodes situated one space diagonally up and to the left of the nodes of B: they form a piece of rim identical in shape to B. The node corresponding to h is not an H (of $\bar{\lambda}$), however, and the node corresponding to f is not an F. Otherwise these q nodes are H's or F's, or both, according as the corresponding nodes of B are H's or F's or both. If there are no nodes preceding B, then there may be less than q new nodes becoming part of the rim of the diagram: the same thing may happen if there are no nodes following B. In any case, however, the above remarks apply to as many new members of the rim as there may be.

Consider first the case where the node removed from λ^*_i is the top node of λ^*_i. When it is removed, all the δ's of λ^*_i are reduced by 1. We must check and see that the hooks of C_i are all reduced by q, and that the other chains remain unaffected as to the lengths of their hooks. B in this case will be the smallest hook of C_i, and h will be the top node of the hooks of C_i. The node corresponding to h (see above) if there is one at all, is not an H, and the node following f, if there is one, becomes an H. Thus we have a chain of $\bar{\lambda}$ whose node is q nodes below h, and the bottom nodes of the hooks remain unaltered. This reduces the lengths of the hooks of C by 1, as required.

Let C_j, $j \neq i$, be a chain with top node c. Then c cannot coincide with h. If it lies above h it is unaffected by the removal of B. (It remains an H.) Those bottom nodes of hooks of C_j which lie below B are unaffected. If a bottom node b of a hook of C_j is a node of B, then it lies at least one column to the right of the column of f, and since c is at least one row above h, and hence above b, b is not in the first row of λ. Hence there is a node corresponding to b—one row above and one column to the left, and this node will be an F. Thus the lengths of the hooks of C_j are unchanged. (More precisely, if λ has a chain C_j, then $\bar{\lambda}$ has a chain of hooks of the same lengths.) Exactly similar arguments deal with the other possible locations of c.

Any chain of $\bar{\lambda}$ corresponds to some chain of λ, for a bottom node of one of its hooks is either an F of λ, corresponding to a node of B, or is the node preceding h on the rim of λ. In the last case the chain corresponds to the chain C_i, and in the other cases it corresponds to the chain of which it (or its corresponding node) is a bottom node. Thus the star diagrams $(\bar{\lambda})^*$ and $\bar{\lambda}^*$ are identical.

Exactly similar arguments deal with the case where the node removed from λ^* is not the top node of λ^*_i, completing the proof of the theorem.

3. Classes of congruent δ's. A diagram is completely determined by its δ's, and these will be our primary concern in the sections to follow. In this section we make some remarks concerning the congruent (mod q) classes of δ's of a right diagram.

Consider the chain, C_i, of kq-hooks corresponding to λ^*_i. Suppose H_i, the top node of these hooks, is the $(m+1)$th node of the rim of λ. The lengths of the hooks are all divisible by q, and their bottom nodes are bottom nodes of δ's of λ. The lengths of these δ's are just m greater than the lengths of the hooks of C_i. Hence they are all congruent (mod q). We shall refer to them as δ^i's.

A δ^i cannot be congruent (mod q) to a δ^j, $i \neq j$, for suppose it were, and suppose $i > j$. Then the bottom node, f, of the δ^i would lie c nodes (along the rim) below the bottom node of the δ^j, where $c \equiv 0 \pmod{q}$, and f would be the bottom node of a hook of C_j. But f is also the bottom node of a hook of C_i, and this cannot be, as was seen in §2. Hence the δ^i's are not congruent to the δ^j's, $i \neq j$.

Since there can be at most q different classes of numbers congruent (mod q), we have immediately

3.1 *The number of constituents of δ^* is at most q.*

This was formerly proven by Robinson, using a theorem of Nakayama's, and by consideration of the removal of hooks from λ.

The following theorem is used in the proof of Theorem **A**.

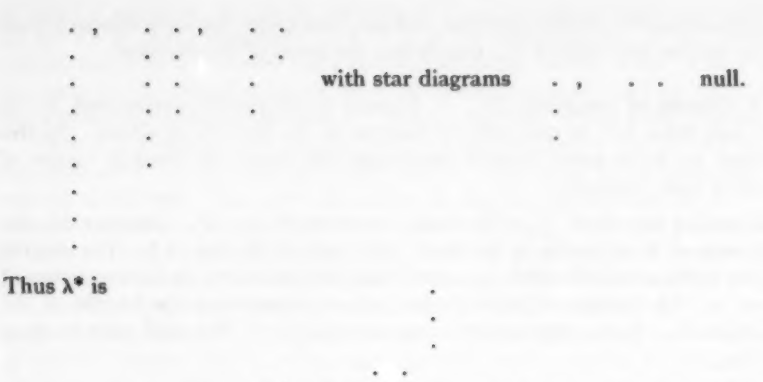
C. *Gather the δ's of λ into classes of δ's which are congruent (mod q). For each such class of congruent δ's form the diagram whose δ's are the δ's of this class. The star diagrams of the diagrams thus formed will be the constituents of λ^*.*

To illustrate the theorem let $q = 3$ and consider the diagram

3.2 λ:

. . . . .
. . . . .
. . . .
. .
.

The δ's are 9, 7, 5, 4, 2, and the classes of congruents δ's are $\{9\}, \{7, 4\}, \{5, 2\}$, which yield the diagrams



Thus λ^* is

To prove **C**, consider a class, K, of congruent δ's, and let μ be the diagram whose δ's are the members of K. We have just seen that the δ's of λ which have bottom nodes in common with hooks of a chain, C_i, are all congruent (mod q). We will show that, if K is the class of all δ's congruent to those associated in this way with some C_i, then $\mu^* = \lambda^*_i$: otherwise μ^* is null.

Suppose that the hooks of C_i, are of lengths k_1q, k_2q, . . . , k_sq, and that their common top node, H_i, is the $(m+1)$th node of the rim of λ. Then the members of K are $k_1q + m$, $k_2q + m$, . . . , $k_sq + m$ and possibly some of $m - q$, $m - 2q$, . . . as well: these will be the δ's of μ. We wish to show that the $(m+1)$th node of the rim of μ is an H, but that the $(m+1-q)$th, $(m+1-2q)$th, . . . nodes are not H's. This will yield a chain of hooks of μ of lengths k_1q, k_2q, . . . , k_sq, as required.

The mth node of the rim μ is not an F, for suppose it were: then the mth node of the rim of λ would be an F also, but this could not be, since the $(m+1)$th node, H_i, is an H. Hence the $(m+1)$th node of the rim of μ is an H.

The $(m+1-bq)$th node, g, of the rim of λ is not an H, for if it were then there would be a chain of kq-hooks beginning at g, and the hooks of C_i would already have been represented in the constituent of λ^* corresponding to this chain, contrary to assumption. Hence the $(m-bq)$th node is an F, and so is the $(m-bq)$th node of the rim of μ. Hence the $(m-bq+1)$th node of the rim of μ is not an H.

Therefore μ has a chain of kq-hooks identical with C_i. It can only have one chain, since all the δ's of μ are congruent (mod q). Hence $\mu^* = \lambda^*_i$.

Finally, suppose that the members of K have no bottom nodes in common with hooks of C_i, for any i. Then μ has no kq-hooks, for, suppose the $(p+1)$th and $(p+kq)$th nodes of its rim are an H and an F respectively. The $(p+kq)$th node of the rim of λ is then an F. The pth node of the rim of μ is not an F, since the $(p+1)$th is an H; hence the pth node of the rim of λ is not an F, and hence the $(p+1)$th node is an H. But this results in a kq-hook of λ

which shares its bottom node with a δ of K, contrary to assumption. Hence μ^* is null.

4. The determination of $e(z_\lambda)$. We shall henceforth require q to be a prime, p. The degree of the irreducible representation corresponding to λ is

$$z_\lambda = n! \frac{\prod_{i<j} (\delta_i - \delta_j)}{\prod_i \delta_i!} \tag{4.1}$$

The degree of the reducible representation corresponding to λ^* is

$$z_{\lambda^*} = \frac{B!}{B_1! \ldots B_k!} \cdot z_{\lambda^*_1} \ldots z_{\lambda^*_k} \tag{4.2}$$

where B_i denotes the number of nodes in λ^*_i, $B = \sum_i B_i$, k is the number of constituents of λ^*, and $z_{\lambda^*_i}$ is given by 4.1 applied to λ^*_i.

We shall prove

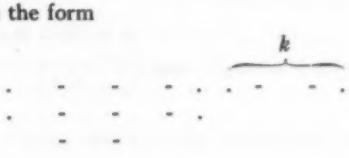
$$\mathbf{A}: \qquad e(z_\lambda) = e(n!) - e((n-A)!) + e(z_{\lambda^*}) \tag{A}$$

by pairing off factors p from z_λ and z_{λ^*}. First we deal with the particular case where (1) λ^* has exactly one constituent, and (2) the kp-hooks of the corresponding chain all begin at the top node of λ—that is, they are all δ's of λ. The general case is reduced to this special case by (1) Theorem C, which reduces the problem to consideration of a single constituent, and (2) Lemma 4.3 which enables us to remove rows from the top of λ until the row at which the kq-hooks begin is reached.

4.3 Lemma. *Suppose λ has no κp-hook beginning in the first row, and suppose the first row of λ is removed, leaving $\bar{\lambda}$ of $\bar{n}$ nodes. Then*

$$e(z_\lambda) - e(z_{\bar{\lambda}}) = e(n!) - e(\bar{n}!).$$

Let us assume λ to have the form



so that the first row is k nodes longer than the second row, where $k \not\equiv 0$. We wish to compare

$$z_\lambda = n! \frac{\prod_{i<j} (\delta_i - \delta_j) \cdot \prod_{f<g} (\delta_f - \delta_g) \cdot \prod_{a<b} (\delta_a - \delta_b)}{\prod_s (\bar{\delta}_s + k + 1)!\, k!\, (k-1)! \ldots 3!\, 2!\, 1}$$

with

$$z_{\bar{\lambda}} = \bar{n}! \frac{\prod_s (\bar{\delta}_s - \bar{\delta}_t)}{\prod_s (\bar{\delta}_s!)},$$

where (a) δ_i, δ_j of λ have their feet (bottom nodes) in $\bar{\lambda}$ (and in λ).

(b) δ_f, δ_g of λ have their feet in the first row of λ. (Hence in the last k nodes of this row.)

(c) δ_a of λ has its foot in $\bar{\lambda}$ (and in λ), δ_b of λ has its foot in the first row of λ.

(d) $\bar{\delta}_s$, $\bar{\delta}_t$ are δ's of $\bar{\lambda}$.

Note that $(\bar{\delta}_s + k + 1)$, as s varies, and k, $(k-1)$, $\ldots$, 3, 2, 1 are just the δ's of λ.

(i) $(\delta_i - \delta_j)$ depends only on the feet of δ_i, δ_j, hence $\prod_{i<j} (\delta_i - \delta_j) = \prod_{s<t} (\bar{\delta}_s - \bar{\delta}_t)$ and $e\big(\prod_{i<j} (\delta_i - \delta_j)\big) = e\big(\prod_{s<t} \bar{\delta}_s - \bar{\delta}_t)\big)$.

(ii) $k < p$, for otherwise there would be a κp-hook in the first row of λ, contrary to assumption. Hence $\delta_f < p$, $\delta_g < p$, and $e\big(\prod_{f<g} (\delta_f - \delta_g)\big)$, $e((k)!\,(k-1)!\,(k-2)!\ldots 3!2!1)$ are both zero.

(iii) $\dfrac{(\bar{\delta}_s + k + 1)!}{\bar{\delta}_s!} = (\bar{\delta}_s + k + 1)(\bar{\delta}_s + k)\ldots(\bar{\delta}_s + 1)$; $(\bar{\delta}_s + k + 1)$ is not divisible by p, since λ was assumed to have no κp-hook beginning in the first row. Hence $e\left(\dfrac{(\bar{\delta}_s + k + 1)!}{\bar{\delta}_s!}\right) = e\big((\bar{\delta}_s + 1)\ldots(\bar{\delta}_s + k)\big)$, and we note that $(\bar{\delta}_s + k), \ldots, (\bar{\delta}_s + 1)$ are the terms of $\prod_{a<b} (\delta_a - \delta_b)$ with a given $\delta_a = \bar{\delta}_s + k + 1$. Hence $e\left[\prod_s \dfrac{(\bar{\delta}_s + k + 1)!}{\prod \bar{\delta}_s!}\right] = e\big(\prod_{a<b} (\delta_a - \delta_b)\big)$. Thus all contributions cancel except those of $n!$ and $\bar{n}!$. These yield the required result.

4.4 LEMMA. *If λ is a p-core, then $e(z_\lambda) = e(n!)$.*

This is proven by removing rows until the last row is reached, and applying 4.3 on the removal of each row.

4.5 COROLLARY. *If λ is a p-core, then* $e\left[\dfrac{\prod_{s<t} (\delta_s - \delta_t)}{\prod_i \delta_i!}\right] = 0.$

The following two lemmas are given without proof.

4.6 LEMMA. $e((pa)!) - e(a!) = a$, *a any integer.*

4.7 LEMMA. *If λ has k columns, then* $\sum_{i=1}^{k} \delta_i = n + \frac{1}{2}k(k-1)$. (A well-known result.[4])

We are now in a position to prove an important special case of the main theorem.

4.8 *If λ^* has exactly one constituent, and each δ of λ^* represents a δ of λ then*

$$e(z_\lambda) = e(n!) - e((n-A)!) + e(z_{\lambda^*}),$$

where z_λ is given by 4.1. We observe that $\delta_s - \delta_t$ makes no contribution to $e(z_\lambda)$ unless $\delta_s \equiv \delta_t \pmod p$. Hence we may write

$$z_\lambda = n! \prod_i \left[\frac{\prod_{s<t} (\delta_s{}^i - \delta_t{}^i)}{\prod_j \delta_j{}^i!} \right] . K$$

where the δ^i's are a class of congruent δ's and $\prod_i$ is taken over these classes, and K makes no contribution to $e(z_\lambda)$. Since λ^* has only one constituent, just one class of δ's yields a diagram which is not a p-core. (See C.) For the other classes, 4.5 tells us that

$$e\left[\frac{\prod_{s<t} (\delta_s{}^i - \delta_t{}^i)}{\prod_j \delta_j{}^i!} \right] = 0.$$

Hence the contributing part of z_λ is just

$$C = n! \frac{\prod_{s<t} (\delta^\circ{}_s - \delta^\circ{}_t)}{\prod_i \delta^\circ{}_i!}$$

where the δ°'s are the δ's corresponding to the δ's of λ^*. Let us denote a δ of λ^* by δ^*. Each δ^* represents a δ of λ, by assumption, and $\delta = p\delta^*$. Hence we may write

$$C = n! \frac{\prod_{s<t} (p\delta^*{}_s - p\delta^*{}_t) \cdot (pB)!}{\prod_i (p\delta^*{}_i)! \quad (pB)!}$$

where the extra unit factor is added for later convenience, and

$$z_{\lambda^*} = B! \frac{\prod_{s<t} (\delta^*{}_s - \delta^*{}_t)}{\prod_i \delta_i{}^*!}. \qquad (B \text{ defined as in 4.2.})$$

It remains to show that

$$e(C) = e(n)! - e(\,(n - A)!) + e(z_{\lambda^*}).$$

By Theorem B′, if a k-hook is removed from λ^*, and the corresponding kq-hook is removed from λ, then the relationship of diagram to star diagram is preserved. This leads directly to the fact that $pB = n - A$, which accounts for the term $e(n!) - e(\,(n - A)!)$. For the remainder, consider:

(1) $$e[\prod_{s<t} (p\delta^*{}_s - p\delta^*{}_t)] - e[\prod_{s<t} (\delta^*{}_s - \delta^*{}_t)].$$

This is seen to be equal to the number of differences $(\delta^*{}_s - \delta^*{}_t)$, which is $\frac{1}{2}\, k(k-1)$, where k is the number of columns of λ^*.

(2) $$e(\,(pB)!) - e(B!) = B, \qquad \text{by 4.6}$$

(3) $$e[\prod_i ((p\delta^*{}_i)!)] - e[\prod_i (\delta^*{}_i!)] = \sum_i \delta^*{}_i \qquad \text{by 4.6}$$
$$= B + \tfrac{1}{2} k(k-1) \qquad \text{by 4.7.}$$

Then (1), (2) and (3) yield the required result.

Next we extend 4.8 by means of 4.3.

4.9 *If λ^* has exactly one constituent, then the result of 4.8 holds.*

Let H be the top node of the kp-hooks of λ which correspond to δ's of λ^*. Remove rows until the row of H is reached. Denote the succession of diagrams obtained by $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_r$, and the number of nodes in λ_i by n_i.

λ_r satisfies the conditions of 4.8, and $\lambda^*_r = \lambda^*$. Suppose the p-core of λ_r has A_r nodes.

$$\begin{aligned} e(z_\lambda) - e(z_{\lambda_r}) &= e(z_\lambda) - e(z_{\lambda_1}) + e(z_{\lambda_1}) - \ldots + e(z_{\lambda_{r-1}}) - e(z_{\lambda_r}) \\ &= e(n!) - e(n_1!) + e(n_1!) - \ldots + e(n_{r-1}!) - e(n_r!), \quad \text{by 4.3} \\ &= e(n!) - e(n_r!). \end{aligned}$$

Since $\lambda^*_r = \lambda^*$, 4.8 yields

$$e(z_{\lambda_r}) = e(n_r!) - e(\,(n_r - A_r)!) + e(z_{\lambda^*}).$$

But $n - A = pB = n_r - A_r$, since $\lambda^*_r = \lambda^*$, hence

$$e(z_\lambda) = e(n!) - e(n_r!) + e(z_{\lambda_r}) = e(n!) - e(\,(n - A)!) + e(z_{\lambda^*}).$$

We can now prove the main theorem.

A: $$e(z_\lambda) = e(n!) - e(\,(n - A)!) + e(z_{\lambda^*}).$$

We have $$z_\lambda = n!\ \prod_i \left(\frac{\prod_{s<t} (\delta_s^i - \delta_j^i)}{\prod_j (\delta_j^i)!} \right) . K$$

where Π_i is taken over classes of congruent δ's and $K \neq 0$. Each class $\{\delta^i\}$ is the class of δ's of a diagram, λ_i, with n_i nodes and a p-core of A_i nodes, say.

If λ_i is a p-core, then, by 4.5

$$e\left(\frac{\prod_{s<t} (\delta_s^i - \delta_t^i)}{\prod_j (\delta_j^i)!} \right) = 0.$$

Hence we need only consider $\prod_i$ to range over classes $\{\delta^i\}$ for which λ_i is not a p-core. We now apply Theorem **C**, proven in **3**. This tells us that the (λ^*_i)'s are just the constituents of λ^*, and we may suppose the λ_i's to be numbered correspondingly. We have

$$z_\lambda = n!\ \prod_i \left(\frac{z_{\lambda_i}}{n_i!} \right) . K,$$

$$\begin{aligned} e\left(\frac{z_{\lambda_i}}{n_i!} \right) &= - e(\,(n_i - A_i)!) + e\,(z_{\lambda^*_i}) \qquad \text{by 4.9} \\ &= - e(\,(pB_i)!) + e(z_{\lambda^*_i}). \end{aligned}$$

Also, (4.2), $$z_{\lambda^*} = \frac{B!}{\prod_i B_i!} \prod_i z_{\lambda^*_i}.$$

Hence $$e(z_\lambda) = e(n!) - \sum_i e((pB_i)!) + \sum_i e(z_{\lambda^{*i}})$$

$$e(z_{\lambda^*}) = e(B!) - \sum_i e(B_i!) + \sum_i (z_{\lambda^{*i}}).$$

Hence $$e(z_\lambda) - e(z_{\lambda^*}) = e(n!) - e(B!) + \sum_i e(B_i!) - e((pB_i)!)$$

$$= e(n!) - e(B!) - \sum_i B_i \qquad \text{by 4.6}$$

$$= e(n!) - e((pB)!) + e((pB)!) - e(B!) - B$$

$$= e(n!) - e((n-A)!) \qquad \text{by 4.6.}$$

5. Star diagrams and the p-series. We have derived Robinson's formula for $e(z_\lambda)$ without making use of the p-series of λ. To make the story complete, however, we should reverse the original procedure carried out by Robinson, and derive the p-series formula for $e(z_\lambda)$ from that which we have just proven. The main task is to find a connection between the star diagram and the p-series.

Let us repeat the operation of forming star diagrams. That is, we form λ^*, then we form the star diagrams of the constituents of λ^*, arrange these in order, forming $(\lambda^*)^*$, and so on. Denote the sequence of diagrams thus formed by $\lambda, \lambda^*, \lambda^{2*}, \ldots, \lambda^{r*}$ and the number of nodes in their p-cores by $A, A_1, A_2, \ldots, A_r$, where λ^{r*} is a p-core. A node removable from λ^{i*} represents a p-hook removable from $\lambda^{(i-1)*}$: continuing back to λ we see that it represents a p^i-hook removable from λ.

Consider the last diagram, λ^{r*}. Each node removable from it represents a p^r-hook of λ, and this must be the longest p^i-hook of λ, for otherwise λ^{r*} would have a p-hook and would not be the last diagram of the sequence. Let us remove a node from λ^{r*}, and the corresponding hooks from $\lambda^{(r-1)*}, \ldots, \lambda^*, \lambda$. The result is the removal of a p^r-hook from λ. When all A_r nodes of λ^{r*} have been removed, we will have removed A_r p^r-hooks from λ, and no more p^r-hooks remain on λ. We will also have removed all nodes from $\lambda^{(r-1)*}$ except its p-core. Removing these one at a time, we remove A_{r-1} p^{r-1}-hooks from what is left of λ. And so on, until all nodes have been removed from λ^*, and only the p-core of λ remains. What we have done is to remove the hooks of the p-series from λ. Hence the p-series of λ is

$$p^r, \ldots (A_r \text{ times}); p^{r-1}, \ldots (A_{r-1} \text{ times}); \ldots \quad ; p, \ldots (A_1 \text{ times})$$

which proves Theorem **D**.

Next we prove Nakayama's formula for $e(z_\lambda)$, namely

5.2 $$e(z_\lambda) = e(n!) - \sum A_i e(p^i!).$$

Let the number of nodes in λ^{i*} be n_i. By removing all the nodes from $\lambda^{(i+1)*}$, and the corresponding p-hooks from λ^{i*} we see that $n_i - A_i = pn_{i+1}$.

5.21 $$n = A_1 + pA_2 + p^2A_3 + \ldots + p^{r-1}A_r.$$

Proof:
$$\begin{aligned} n_1 - A_1 &= pn_2 \\ &= p(A_2 + pn_3) \\ &= pA_2 + p^2(A_3 + pn_4) \\ &= pA_2 + p^2A_3 + p^3A_4 + \ldots + p^{r-1}A_r \end{aligned}$$

by a simple induction.

The proof of 5.2 is based on an induction over n. Since each B_i is less than n, we may assume the theorem for each of the constituents of λ^*. The p-series for λ^* is just the sum of the p-series of its constituents, and hence we may assume the theorem true for λ^*. The p-series of λ^* is

$$p, \ldots (A_2 \text{ times}); p^2, \ldots (A_3 \text{ times}); \ldots ; p^r \ldots (A_{r+1} \text{ times}).$$

Hence induction over n implies that

$$e(z_{\lambda^*}) = e(n_1!) - [A_2e(p!) + A_3e(p^2!) + \ldots + A_re(p^{r-1}!)].$$

By Theorem **A**, $e(z_\lambda) = e(n!) - e\,((n - A)!) + e(z_{\lambda^*})$
$= e(n!) - e\,((pn_1)!) + e(z_{\lambda^*})$.

Using the above expression for $e(z_{\lambda^*})$, we have

$$\begin{aligned} e(z_\lambda) &= e(n!) - e(\,(pn_1)!) + e(n_1!) - [A_2e(p!) + A_3e(p^2!) + \ldots + A_re(p^{r-1}!)] \\ &= e(n!) - n_1 - [A_2e(p!) + \ldots + A_re(p^{r-1}!)] \\ &= e(n!) - [A_1 + pA_2 + \ldots + p^{r-1}A_r] \qquad \text{by 5.21} \\ &\quad - [A_2e(p!) + \ldots + A_re(p^{r-1}!)] \\ &= e(n!) - [A_1 + A_2(e(p^2!) - e(p!) + \ldots] \\ &\quad - [A_2e(p!) + \ldots + A_r\,e(p^{r-1}!)\,] \\ &= e(n!) - [A_1e(p!) + A_2e(p^2!) + \ldots + A_re(p^r!)]. \end{aligned}$$

The theorem is easily established for $n = 1$, or other small values, which is sufficient to start the induction.

References

[1] Brauer, R. and Nesbitt, C., *Ann. of Math.*, vol. 42 (1941), 556-590.
[2] ——— and Robinson, G. de B., *Trans. Roy. Soc. Can.*, vol. XLI, series III, sec. III (1947), 11-25.
[3] Nakayama, T., *Jap. J. Math.*, vol. 17 (1941), 411-423.
[4] ——— *Jap. J. Math.*, vol. 17 (1940), 165-184.
[5] Robinson, G. de B., *Amer. J. Math.*, vol. LXX,(1948), 277-294.
[6] ——— *Amer. J. Math.*, vol. LXIX, (1947), 286-298.

University of New Brunswick

COMBINATORIAL PROBLEMS

S. CHOWLA AND H. J. RYSER

1. Introduction. Let it be required to arrange v elements into v sets such that every set contains exactly k distinct elements and such that every pair of sets has exactly $\lambda = k(k-1)/(v-1)$ elements in common $(0 < \lambda < k < v)$. This combinatorial problem is studied in conjunction with several similar problems, and these problems are proved impossible for an infinitude of v and k. An incidence matrix is associated with each of the combinatorial problems, and the problems are then studied almost entirely in terms of their incidence matrices. The techniques used are similar to those developed by Bruck and Ryser for finite projective planes [3]. The results obtained are of significance in the study of Hadamard matrices [6; 8], finite projective planes [9], symmetrical balanced incomplete block designs [2; 5], and difference sets [7].

2. Combinatorial problems. Let $x_1, x_2, \ldots, x_v$ denote v elements and let $s_1, s_2, \ldots, s_v$ denote v sets formed from these elements. Let the elements $x_1, x_2, \ldots, x_v$ be listed in a row and let the sets $s_1, s_2, \ldots, s_v$ be listed in a column. Let 1 be inserted in row i and column j if the element x_j belongs to the set s_i, and 0 in the contrary case. The matrix A of order v formed from this square array of zeros and ones is called the *incidence matrix* of the arrangement of v elements into v sets. Clearly the incidence matrix serves to characterize this arrangement completely. We proceed now to consider a series of combinatorial problems, and to study these problems in terms of their incidence matrices.

PROBLEM I. *Arrange v elements into v sets such that*

(I_1) *every set contains exactly k distinct elements,*

(I_2) *every pair of sets has exactly $\lambda = k(k-1)/(v-1)$ elements in common $(0 < \lambda < k < v)$.*

PROBLEM II. *Arrange v elements into v sets such that*

(II_1) *each element occurs in exactly k distinct sets,*

(II_2) *every pair of elements occurs in the v sets exactly $\lambda = k(k-1)/(v-1)$ times $(0 < \lambda < k < v)$.*

PROBLEM II′. *Arrange v elements into v sets fulfilling* (II_1) (II_2), *and* (I_1).

PROBLEM III. *Arrange v elements into v sets fulfilling* (I_1), (I_2), (II_1), *and* (II_2).

Received January 6, 1949.

PROBLEM IV. *Arrange v elements into v sets fulfilling* (I_1) *and* (I_2) *in such a way that the incidence matrix of the arrangement is cyclic, i.e.*

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_v \\ a_2 & a_3 & \dots & a_1 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_v & a_1 & \dots & a_{v-1} \end{bmatrix}.$$

Problem I has a solution if and only if there exists a matrix A of order v composed of zeros and ones such that

$$A A^T = B, \tag{I}$$

where A^T denotes the transposed matrix of A and B is a symmetric matrix with k in the main diagonal and λ in all other positions. Problem II requires

$$A^T A = B, \tag{II}$$

and Problem III requires

$$AA^T = A^T A = B. \tag{III}$$

The preceding problems arise naturally in certain combinatorial investigations. Problem I for $v = 4n - 1$, $k = 2n - 1$, and $\lambda = n - 1$ was proposed by Todd, and was shown to be equivalent to finding a Hadamard matrix of order $4n$ [6; 8]. Problem II′ was studied by Bose, and the arrangements obtained were called symmetrical balanced incomplete block designs [2; 5]. Veblen and Bussey introduced the finite projective plane, and Problem III for $v = N^2+N+1$, $k = N + 1$, $N \geqslant 2$, and $\lambda = 1$ is equivalent to finding a projective plane with $N + 1$ points on a line [3; 9].

Singer defined a difference set of k numbers mod v as a set of integers d_1, $d_2, \dots, d_k$ such that the congruences $d_i - d_j \equiv n \bmod v$ have the same number of solutions $\lambda = k(k-1)/(v-1)$ for every $n \not\equiv 0 \bmod v$ [7]. Problem IV is equivalent to finding a difference set of k numbers mod v. For if such a difference set exists, form the array of k rows and v columns

$$\begin{matrix} d_1, d_1 - 1, \dots, d_1 - (v-1) \\ \cdot \\ \cdot \\ \cdot \\ d_k, d_k - 1, \dots, d_k - (v-1) \end{matrix}$$

where the integers are reduced mod v so that they lie in the range $1 \leqslant x \leqslant v$. Now form an incidence matrix A of order v by taking column i of the above array and placing in row i of the matrix A ones in columns $d_1 - (i-1)$, $\dots, d_k - (i-1)$ and zeros in all other positions. Clearly, A by the nature of its construction is cyclic. Moreover, A has exactly k ones in each row, and since for $r \not\equiv s$, $d_i - r \equiv d_j - s \bmod v$ has exactly λ solutions, any two rows of A have exactly λ ones in common. Thus the matrix A yields a solution of Problem IV. Conversely, suppose that Problem IV has a solution. Then the

first row of the incidence matrix A has ones in the k columns $d_1, \ldots, d_k$, and these k numbers form a difference set mod v. For row $n+1$ of A has ones in the columns $d_1 - n, \ldots, d_k - n$, where the integers are taken mod v, and for $n \not\equiv 0 \bmod v$, the sets $d_1, \ldots, d_k$ and $d_1 - n, \ldots, d_k - n$ have exactly λ elements in common. Hence $d_i - d_j \equiv n \bmod v$ has exactly λ solutions.

3. Identical combinatorial problems. Let P and Q be any two of the preceding combinatorial problems. The problems P and Q are said to be identical, written $P \equiv Q$, provided that each solution of P is necessarily a solution of Q, and conversely each solution of Q is necessarily a solution of P.

THEOREM 1. *Problem* I $\equiv$ *Problem* II $\equiv$ *Problem* II$'\equiv$ *Problem* III.

Suppose that A is a matrix of order v composed of zeros and ones such that $AA^T = B$, where B has k in the main diagonal and $\lambda = k(k-1)/(v-1)$ in all other positions. For this A we prove that $A^TA = B$. Define the matrix O of order $v+1$ by the equation

$$O = \begin{bmatrix} -k & \sqrt{-\lambda} & \cdots & \sqrt{-\lambda} \\ \sqrt{-\lambda} & & & \\ \cdot & & A & \\ \cdot & & & \\ \cdot & & & \\ \sqrt{-\lambda} & & & \end{bmatrix}.$$

Recalling that $\lambda = k(k-1)/(v-1)$, an easy computation shows that $OO^T = (k-\lambda)I$, where I is the identity matrix of order $v+1$. But then $OO^T = O^TO$, and then by the very structure of O, it follows that $AA^T = A^TA$.

Thus a solution of Problem I is necessarily a solution of Problem III, and consequently Problem I $\equiv$ Problem III. Moreover, the matric equation $A^T(A^T)^T = A^TA = B$ now implies that $AA^T = B$, and consequently Problem I $\equiv$ Problem II. This proves Theorem I. (For another proof see Bose [2].)

THEOREM 2. *There exist values for v and k for which Problem* III *has a solution and for which Problem* IV *has no solution.*

Evidently every solution of Problem IV is a solution of Problem III. To prove Theorem 2 we utilize the following theorem of Chowla, which establishes the nonexistence of a certain class of difference sets. The recent investigations of Marshall Hall have also been successful in proving the nonexistence of large classes of such sets [4].

Let $v, k,$ and $\lambda = k(k-1)/(v-1)$ be positive integers, $0 < \lambda < k < v$. Let $p \equiv 3 \bmod 4$ be a prime factor of v and let q be an odd prime factor which divides the squarefree part of $k-\lambda$. If the Legendre symbol $(-p|q) = -1$, then there does not exist a difference set of k numbers mod v.

To prove the theorem let $d_1, d_2, \ldots, d_k$ denote such a difference set, and define $S = \sum_{j=1}^{k} \rho^{d_j}$, where $\rho = e^{\frac{2\pi i}{p}}$. Then $S\overline{S} = k + \lambda(\rho + \rho^2 + \ldots + \rho^{v-1})$

$= k - \lambda$, where $\bar{S}$ denotes the complex conjugate of S. Let $t = S\theta^2 S\theta^4 S \ldots \theta^{p-3} S$, where θ denotes a generating automorphism of the cyclic algebraic field $R(\rho)$. If $N(S)$ denotes the algebraic norm of S in $R(\rho)$, then $N(S) = t\theta t$. The algebraic integers t and θt are conjugates in the unique quadratic subfield $R\left(\sqrt{(-1)^{\frac{p-1}{2}} p}\right)$ of $R(\rho)$ [1; 10]. Consequently $N(S) = (x^2 + py^2)/4$, where x and y are integers. But the equation $S\bar{S} = (k - \lambda)$ implies $N(S) = (k - \lambda)^{\frac{p-1}{2}}$. Thus $x^2 + py^2 - 4(k - \lambda)^{\frac{p-1}{2}} = 0$, and this equation may be rewritten in the form $x^2 + py^2 - qtz^2 = 0$, where t is squarefree and prime to q, and where x, y, and z do not have a prime factor in common. It now follows that q does not divide y, and hence $(y^{-1}x)^2 \equiv -p \bmod q$.

Now let $v = 55$ and $k = 27$. Then $\lambda = 13$ and $k - \lambda = 14$. Select $p = 11$ and $q = 7$. Then $(-11|7) = -1$, and consequently there does not exist a difference set of 27 numbers mod 55. Thus for these values of v and k, Problem IV does not have a solution. On the other hand it is well known that a Hadamard matrix of order 56 exists, and by the remarks of Todd, Problem I has a solution for these values of v and k [6; 8]. But Problem I $\equiv$ Problem III.

4. The impossibility of certain combinatorial problems. In this section the impossibility of Problem I is proved for an infinitude of v and k. Clearly the impossibility of Problem I for a given v and k implies the impossibility for the same v and k of Problems II, II′, III, and IV. Interpreted with regard to the results of the previous sections, the theorems which follow offer generalizations of numerous previous investigations. Actually Theorems 4 and 5 are rather straightforward generalizations of a theorem of Bruck and Ryser on the nonexistence of certain finite projective planes, and for projective planes these theorems give no new information [3]. However, their proofs are independent of the difficult Minkowski-Hasse theory of the invariants of a rational quadratic form under rational transformations. (The writers are indebted to Daniel Zelinsky for helpful comments concerning the proof of Theorem 5.)

THEOREM 3. *If v is even and if $k - \lambda$ is not a square, then Problem* I *has no solution.*

A solution of Problem I implies that $AA^T = B$, where B has k in the main diagonal and λ in all other positions. Subtract column one of the matrix B from each of the other columns, and then add to row one each of the other rows. It readily follows that the determinant of B is given by

$$\det B = \det^2 A = (k - \lambda)^{v-1}(k + (v-1)\lambda) = (k - \lambda)^{v-1} k^2.$$

Thus if v is even and if $k - \lambda$ is not a square, then Problem I has no solution.

THEOREM 4. *If $v \equiv 1 \bmod 4$ and if there exists an odd prime p such that p divides the squarefree part of $k - \lambda$, and, moreover, if $(\lambda|p) = -1$, then Problem* I *has no solution.*

The matric equation $AA^T = B$ implies

$$k\sum_{i=1}^{v} x_i^2 + \lambda \sum_{i \neq j} x_i x_j = (k-\lambda)\sum_{i=1}^{v} x_i^2 + \lambda(\sum x_i)^2 = \sum_{i=1}^{v} u_i^2,$$

where the matrix $C = [c_{ij}]$ of the transformation $x_i = \sum c_{ij}u_j$ is rational and nonsingular. By the four-square theorem of Lagrange, $k - \lambda = a_1^2 + a_2^2 + a_3^2 + a_4^2$, where the a's are integers. If

$$A = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & -a_1 & a_4 & -a_3 \\ -a_3 & a_4 & a_1 & -a_2 \\ a_4 & a_3 & -a_2 & -a_1 \end{bmatrix},$$

then $AA^T = (k-\lambda)I$, where I is the identity matrix of order 4. Thus if $[k-\lambda, k-\lambda, \ldots, k-\lambda]$ is a diagonal matrix of order $v \equiv 1 \bmod 4$, then there exists a rational and nonsingular D such that

$$[k-\lambda, k-\lambda, \ldots, k-\lambda] = D^T[1, 1, \ldots, 1, k-\lambda]\, D,$$

whence $(k-\lambda)\sum_{i=1}^{v} x_i^2 = \sum_{i=1}^{v-1} y_i^2 + (k-\lambda)y_v^2$. Thus

$$\sum_{i=1}^{v-1} y_i^2 + (k-\lambda)y_v^2 + \lambda(\sum d_i y_i)^2 = \sum_{i=1}^{v} u_i^2,$$

where the d_i are rational and the matrix $E = [e_{ij}]$ of the transformation $y_i = \sum e_{ij}u_j$ is rational and nonsingular.

Now set $y_1 = \sum e_{1j}u_j = \pm u_1$, where the coefficient is $+1$ if $e_{11} \neq 1$ and -1 if $e_{11} = 1$. Then $y_2 = \sum_{j=2}^{v} f_j u_j$, and set $y_2 = \pm u_2$, where the coefficient is $+1$ if $f_2 \neq 1$ and -1 if $f_2 = 1$. Continue the process inductively until $y_{v-1} = g_{v-1}u_{v-1} + g_v u_v$, where $y_{v-1} = \pm u_{v-1}$. Now let u_v equal a nonzero rational. Then $u_1, \ldots, u_{v-1}$ are uniquely determined, and, moreover, $y_i = \pm u_i$, for $i = 1, 2, \ldots, v-1$. Thus the Diophantine equation

$$x^2 = (k-\lambda)y^2 + \lambda z^2$$

has a solution in integers other than the zero solution. The equation may be rewritten in the form

$$x^2 = pty^2 + \lambda z^2,$$

where t is squarefree and prime to p, and where x, y, and z do not have a prime factor in common. Now p does not divide z, and hence $(z^{-1}x)^2 \equiv \lambda \bmod p$.

THEOREM 5. *If $v \equiv 3 \bmod 4$ and if there exists an odd prime p such that p divides the squarefree part of $k-\lambda$, and, moreover, if $(-\lambda|p) = -1$, then Problem I has no solution.*

Suppose that $v \equiv 3 \bmod 4$ and that $B = AA^T$. Then

$$\begin{bmatrix} & & 0 \\ & & 0 \\ & B & \cdot \\ & & \cdot \\ & & \cdot \\ & & 0 \\ 0\,0\,\ldots\,0 & & k-\lambda \end{bmatrix} = \begin{bmatrix} & & 0 \\ & & 0 \\ & A & \cdot \\ & & \cdot \\ & & \cdot \\ & & 0 \\ 0\,0\,\ldots\,0 & & 1 \end{bmatrix} [1, 1, \ldots, 1, k-\lambda] \begin{bmatrix} & & 0 \\ & & 0 \\ & A^T & \cdot \\ & & \cdot \\ & & \cdot \\ & & 0 \\ 0\,0\,\ldots\,0 & & 1 \end{bmatrix},$$

and

$$k\sum_{i=1}^{v} x_i^2 + (k-\lambda)x_{v+1}^2 + \lambda \sum_{\substack{i,j=1 \\ i \neq j}}^{v} x_i x_j = (k-\lambda)\sum_{i=1}^{v+1} x_i^2 + \lambda\Big(\sum_{i=1}^{v} x_i\Big)^2$$

$$= \sum_{i=1}^{v} u_i^2 + (k-\lambda)\,u_{v+1}^2,$$

where the matrix $C = [c_{ij}]$ of the transformation $x_i = \sum c_{ij}u_j$ is rational and nonsingular. If $[k-\lambda, k-\lambda, \ldots, k-\lambda]$ is a diagonal matrix of order $v+1 \equiv 0 \bmod 4$, then there exists a rational and nonsingular D such that

$$[k-\lambda, k-\lambda, \ldots, k-\lambda] = D^T[1, 1, \ldots, 1]D,$$

whence $(k-\lambda)\sum_{i=1}^{v+1} x_i^2 = \sum_{i=1}^{v+1} y_i^2$. Thus

$$\sum_{i=1}^{v+1} y_i^2 + \lambda(\sum d_i y_i)^2 = \sum_{i=1}^{v} u_i^2 + (k-\lambda)\,u_{v+1}^2,$$

where the d_i are rational and the matrix $E = [e_{ij}]$ of the transformation $y_i = \sum e_{ij}u_j$ is rational and nonsingular.

Now set $y_1 = \sum e_{1j}u_j = \pm u_1$, where the coefficient is $+1$ if $e_{11} \neq 1$ and -1 if $e_{11} = 1$. Then $y_2 = \sum_{j=2}^{v+1} f_j u_j$, and set $y_2 = \pm u_2$. Continue inductively until $y_v = g_v u_v + g_{v+1}u_{v+1}$, where $y_v = \pm u_v$. Now let u_{v+1} equal a nonzero rational. Then $u_1, \ldots, u_v$ are uniquely determined and $u_i = \pm y_i$ for $i = 1, 2, \ldots, v$. Thus

$$x^2 + \lambda y^2 = (k-\lambda)z^2$$

has a solution in integers other than the zero solution. It follows that $-\lambda$ is a quadratic residue of p. This completes the proof of Theorem 5.

Theorems 4 and 5 may also be derived by the methods employed in [3]. Only minor modifications in the proof given for projective planes are required. It can be shown that for v odd, the matric equation $AA^T = B$ is possible for a rational and nonsingular A if and only if

$$(k-\lambda, -1)_p^{\frac{(v-1)v}{2}}(k-\lambda, v)_p = +1$$

for every odd prime p. The notation $(m, n)_p$ designates the norm-residue symbol of Hilbert. It is easy to verify that this condition excludes precisely those values of v and k covered by Theorems 4 and 5.

References

[1] Bachmann, Paul, *Die Lehre von der Kreistheilung* (Leipzig, 1872), 204.
[2] Bose, R. C., "On the Construction of Balanced Incomplete Block Designs," *Annals of Eugenics*, vol. 9 (1939), 353-399.
[3] Bruck, R. H. and Ryser, H. J., "The Nonexistence of Certain Finite Projective Planes," *Can. J. Math.*, vol. 1 (1949), 88-93.
[4] Hall, Marshall, "Cyclic Projective Planes," *Duke Math. J.* (1947), 1079-1090.
[5] Levi, F. W., *Finite Geometrical Systems* (University of Calcutta, 1942).
[6] Paley, R. E. A. C., "On Orthogonal Matrices," *J. of Math. and Physics*, vol. 12 (1933), 311-320.
[7] Singer, James, "A Theorem in Finite Projective Geometry and Some Applications to Number Theory," *Trans. Amer. Math. Soc.*, vol. 43 (1938) 377-385.
[8] Todd, J. A., "A Combinatorial Problem," *J. of Math. and Physics*, vol. 12 (1933), 321-333.
[9] Veblen, O. and Bussey, W. H., "Finite Projective Geometries," *Trans. Amer. Math. Soc.*, vol. 7 (1906), 241-259.
[10] van der Waerden, B. L., *Moderne Algebra*, I (Berlin, 1937), 165.

University of Kansas
and
Ohio State University

SPHERICAL GEOMETRIES AND MULTIGROUPS

WALTER PRENOWITZ

1. Introduction. The notion *spherical geometry* is suggested by the familiar geometry of the Euclidean 2-sphere in which the role of path is played by "arc of great circle". The first postulational treatment of the subject seems to be that of Halsted [10] for the two-dimensional case. Kline [11] under the name *double elliptic geometry*, gave a greatly simplified foundation for the three-dimensional case based on the primitive notions *point* and *order*.[1] Halsted and Kline study not merely descriptive (that is positional, non-metrical) properties of figures but also introduce metrical notions by postulating or defining congruence. Kline includes a continuity postulate designed to yield real spherical geometry.

Our object is to study the descriptive properties of spherical geometries by general mathematical methods under the weakest possible hypotheses. Just as there exist affine or projective geometries of arbitrary dimension corresponding to any coefficient field (not necessarily commutative), we should like to define spherical geometries of arbitrary dimension corresponding to any *ordered* field. This is possible if we consider the prototype of a spherical geometry to be the set of rays emanating from a point of an ordered affine geometry. This model of course is suggested by the familiar isomorphic mapping of a Euclidean 2-sphere into the family of rays which emanate from its centre. The model does not always (that is for all underlying fields) enjoy all the metrical properties of a Euclidean sphere, but it does exhibit the familiar descriptive properties and it does yield, in a sense, a "topological" sphere for every ordered field.

In order to give spherical geometries an autonomous existence we characterize them abstractly by postulates taking *point* as primitive. To do justice to ordinary goemetrical intuition we follow Kline in adopting as the second primitive notion the 3-term relation *order* suggested by the relation of points a,b,c when b is interior to the minor arc of a great circle which joins a and c. However this relation, despite its intuitive salience, does not facilitate generalization—it is, so to speak, too strongly linear or one-dimensional. Thus we define from it the notion *join* of a pair of points, which can be generalized to sets and extended to n points and forms the basis of our treatment of spherical geometries.

Consider then the following postulates involving a set S of elements $a,b,c,\ldots$ called *points* and a 3-term relation *order* indicated (abc), which may be read *points a,b,c are in the order abc, or b lies between a and c*:

Received February 18, 1949.

[1] Hallett [9], Flanders [8] have also given treatments of the subject.

O1. *If* (abc) *then* a,b,c *are distinct.*

O2. *If* (abc) *then* (cba).

O3. *For each point* a *there exists a unique point* p *such that* $p \neq a$ *and* (axy) *always implies* (xyp).

DEFINITION 1. The uniquely determined point p in O3 is called the *opposite* of a and is denoted functionally by a'.

O4. *If* $b \neq a,a'$ *there exists* x *such that* (axb).

DEFINITION 2. If a,b are points and $b \neq a,a'$ the set of all x for which (axb) is called the *join* or *sum* of a,b and is denoted operationally by $a + b$. The join of a and a, denoted $a + a$, we take to consist of a itself. For simplicity we shall identify element a and set (a) whose only member is a,[2] so that for example we may assert the idempotent law $a + a = a$. This also enables us to employ the inclusion signs $\subset$, $\supset$ for elements as well as sets. For the present we do not define $a + a'$ (see sec. 3).

In order to iterate the operation $+$ we must define sum of *sets* of points. Thus we introduce

DEFINITION 3. If A,B are non-void sets of points $A + B$, the *join* or *sum* of A and B is the set union $\sum_{a \subset A,\ b \subset B} (a + b)$. Observe that this is consistent with the definition of join of *points* a,b just adopted since if A,B consist of single elements, say $A = a$, $B = b$ then $A + B$ as defined reduces to $a + b$.[3] Further note that if any element of B is the opposite of an element of A then $A + B$ is meaningless, since one of the summands in its definition is not significant.

O5. $(a + b) + c = a + (b + c)$ provided both members are defined.

Observe that O5 involves the restrictions $b \neq a'$, that c shall not be the opposite of any point of $a + b$, etc. We shall consider later (sec. 3) the matter of removing these restrictions.

Now we formally define the sense in which the term spherical geometry is to be employed.

DEFINITION 4. A set S in which is defined a relation (abc) satisfying O1, . . . , O5 is called a *spherical geometry.*

If we take S to be a Euclidean n-sphere and (abc) to mean b is an interior point of the minor arc of a great circle which joins a and c then O1, . . . , O5 are satisfied, and we call S with order thus defined a *Euclidean* spherical geometry. A second type of spherical geometry (which includes the first in the sense of isomorphism) arises if S is the set of rays emanating from a point P of an ordered affine space of arbitrary (finite or infinite) dimension and (abc) means ray b is interior to the angle formed by the non-opposite rays a,c. We can form an analogous class of *analytic* spherical geometries as follows. In a

[2] In virtue of this agreement which is maintained throughout the paper (whether or not the elements are points) our definitions and theorems concerning non-void sets hold also for elements.

[3] A similar consistency principle holds throughout the paper whenever a notion defined for sets is apparently ambiguous when applied to elements.

linear vector space L over an ordered division ring F, define $[\alpha\ \beta\ \gamma]$ to mean there exist X, Y in F satisfying

$$\beta = X\alpha + Y\gamma, \qquad X + Y = 1, \qquad 0 < X, Y.$$

Define $\vec{\alpha}$, the *ray* of L determined by α a non-zero element of L, to be the set of $X\alpha$ where $X > 0$ and ranges over F. Call rays $\vec{\alpha}$, $-\vec{\alpha}$ *opposite*. Then S, the set of all rays of L, is a spherical geometry if we define (abc) to mean rays a,c are distinct and not opposite and $a = \vec{\alpha}$, $b = \vec{\beta}$, $c = \vec{\gamma}$ where $[\alpha\beta\gamma]$. It is noteworthy that any spherical geometry of sufficiently high "dimension" is representable as such an *analytic* spherical geometry just as any projective geometry of dimension greater than 2 can be coordinatized—the proof of the former can be made to depend on the latter and will be given elsewhere.

The postulate set O1, . . . , O5 was evolved from that of Kline [11, Axioms I, . . . , X] with the object of formulating a simple and natural basis for spherical geometry and facilitating the study of the operation join. In O5 the only essential novelty is an *operational* formulation of a triangle transversal postulate used by Kline [11, Axiom VII][4] generalized so as to include as many degenerate cases as possible within the limits imposed by the restriction in our definition of $+$; this gives it increased deductive power since it covers linear as well as two-dimensional cases.

Our procedure in the study of spherical geometries will be to exploit consistently the algebraic properties of the operation join. We shall show that by adjoining an ideal element o to spherical geometry S to play the role of an identity element and by extending the definition of $+$ appropriately we can convert S into a generalized group with many-valued composition, called a *multigroup*.[5] The multigroups thus generated are in a class which we call *regular* multigroups; these bear close analogies to abelian groups since each element a has a unique inverse $-a$ and subtraction is related to addition by the familiar formula, $a - b = a + (-b)$. Thus we are able to subsume the theory of spherical geometries under that of regular multigroups, in fact we show it is equivalent in a certain sense to the theory of a particular class of regular multigroups (Theorems 12, 13).

It is known [14, 15] that projective and descriptive (ordered linear) geometries can be characterized and developed as multigroups,[6] which however do not bear close *formal* analogies to abelian groups or to the multigroups which have received most attention from algebraists. On the other hand our regular multigroups are covered by the multigroup theory of Dresher and Ore[6] which

[4] Compare Veblen [17, Assumption 5]; see also Flanders [8, Axiom 05] and his reference to Hallett [9].

[5] A multigroup is a system closed under an associative many-valued operation $\circ$, which contains elements x,y satisfying the relations $a \circ x \supset b$, $y \circ a \supset b$ when a,b are in the system; see [6, pp. 706, 707]. For references on multigroups see J. E. Eaton, "Associative multiplicative systems," *Amer. J. Math.*, vol. 62 (1940) 222-232; also see J. Kuntzmann, "Contribution à l'étude des systèmes multiformes," *Ann. Sciences Toulouse*, (4) vol. 3 (1939), 155-194.

[6] For a "simultaneous" formulation of these geometries as multigroups see [16].

was motivated by algebraic considerations suggested by group theory. Thus from the viewpoint of this paper it would seem that spherical geometries are more "regular" than descriptive or projective geometries and may possibly deserve a more central position in the comparative theory of geometrical systems.

We explicitly develop the theory of spherical geometries only to the point necessary to establish their equivalence to a class of regular multigroups. We then outline the theory of regular multigroups including: subsystems and their generation; cosets, homomorphisms and factor multigroups; linear independence and rank. These ideas cover the geometrical topics: linear (or spherical) subspaces, their alignment and intersection properties; half-spaces (for example in a Euclidean spherical geometry, semicircles, hemispheres, etc.); separation of linear subspaces; linear independence and dimension. For the sake of concreteness and familiarity we use Euclidean spherical geometries to exhibit the geometrical significance of the above algebraic ideas, although they are applicable to arbitrary spherical geometries with no essential change in the discussion.

2. Order properties. In this section we develop the theory of order in a spherical geometry S from postulates O1, . . . , O5 to prepare for the extension of the associative law for $+$. The main results (Theorems 7,8) give combinatory formulas for certain sums of points. Theorems 1, . . . , 6 are principally theorems or postulates of Kline [11] and are intuitively very familiar.

THEOREM 1. *(axa') and $(aa'x)$ are always false.*[7]

Proof. By O3, (axa') implies $(xa'a')$ which is contrary to O1. Similarly $(aa'x)$ implies $(a'xa')$ contrary to O1.

COROLLARY. $(a')' = a$; or equivalently $b = a'$ implies $a = b'$.

Proof. Let a'' denote $(a')'$. Suppose $a'' \neq a$. By O3 $a'' \neq a'$. Thus by O4 (axa'') holds for some x. Thus by O3 $(xa''a')$, which by O2 implies $(a'a''x)$. This contradicts Theorem 1, so that $a'' = a$.

The following result enables us to interpret order relations in "join" language and vice versa.

THEOREM 2. *(abc) implies $b \subset a + c$; conversely $b \subset a + c$ implies (abc) provided $a \neq c$.*

Proof. Suppose (abc). Then $c \neq a$ by O1 and $c \neq a'$ by Theorem 1. Thus $b \subset a + c$ by Definition 2. The remainder of the theorem is immediate by Definition 2.

Next we prove (Kline [11, Axiom V])

THEOREM 3. *(abc), (acd) imply (abd).*

Proof. Suppose (abc), (acd). By Theorem 2 we have

$$b \subset a + c, \qquad c \subset a + d. \tag{1}$$

We wish to assert

$$b \subset a + (a + d) = (a + a) + d. \tag{2}$$

[7] Compare Kline [11, Axiom I, Theorem 3].

The first relation in (2) is implied by (1) in view of Definition 3, provided the expression $a + (a + d)$ is significant; that is, provided $a' \not\subset a + d$. Suppose $a' \subset a + d$. By O1, (acd) implies $a \neq d$. Thus the second part of Theorem 2 implies $(aa'd)$, contrary to Theorem 1. Thus the first relation in (2) is justified. By O5 the second relation in (2) is valid, provided the expression $(a + a) + d$ is significant. But by the idempotent law $(a + a) + d$ reduces to $a + d$ whose existence is involved in relation (1). Thus (2) is verified and it implies $b \subset a + d$. By the second part of Theorem 2 we have (abd) and the proof is complete.

COROLLARY 1. (abc) implies that (bca) is false.

Proof. Suppose (abc), (bca). Then (acb), which with (abc) implies by Theorem 3 (abb).

COROLLARY 2. (abc), (acd) imply (bcd).

Proof. By O2, O3 we have the following implications: $(acd) \to (dca) \to (cad')$. Also $(abc) \to (cba)$. By Theorem 3, O2, O3 and the corollary to Theorem 1 we have

$$(cba), (cad') \to (cbd') \to (d'bc) \to (bcd).$$

We continue with three theorems on order of four points [11, Theorems 22, 24, 25]. We dispense with the proofs—the first depends on the associative law like Theorem 3 and the latter two then follow by standard arguments of the foundations of geometry.

THEOREM 4. *If* (abc), (bcd) *and* $d \neq a'$ *then* (abd) *or* $(ab'd)$.

THEOREM 5. *If* (abc), (abd) *and* $c \neq d$ *then* (acd) *or* (adc).

THEOREM 6. *If* (axb), (ayb), $x \neq y$ *then* (axy) *or* (ayx).

We now prove the principal results of this section.

THEOREM 7. $a + (a' + b) = a + b \cup b \cup a' + b$ *provided*[8] $b \neq a, a'$.

Proof. Suppose $b \neq a, a'$. Then $a' + b$ is defined and is non-void. Furthermore $a' \subset a' + b$ implies $(a'a'b)$ since $a' \neq b$. Thus $a' \not\subset a' + b$ and $a + (a' + b)$ is significant. Clearly the right member of the relation to be established is significant. We shall complete the proof by showing the equivalence of the following relations:

(1) $$x \subset a + (a' + b),$$

(2) $$x \subset a + b \cup b \cup a' + b.$$

Suppose (1). Then by Definition 3

(3) $$x \subset a + y, \qquad y \subset a' + b$$

holds for some y. The second relation in (3) implies $(a'yb)$. From this we have (yba) and so (aby). Thus $a \neq y$ and the first relation in (3) implies (axy). If $x = b$ then (2) holds. Suppose $x \neq b$. Then by Theorem 6 (axy), (aby) imply (axb) or (abx). If (axb) then $x \subset a + b$. If (abx) then (bxa') so that $(a'xb)$ and $x \subset a' + b$. In either case (2) holds.

[8] We use the symbol $\cup$ to denote set theoretic addition. In expressions involving $+$, $\cup$ we adopt the convention that portions separated by $\cup$ signs are to be considered enclosed in parentheses unless the contrary is explicitly indicated.

Conversely we show (2) implies (1). First suppose $x \subset a + b$. Then (axb) since $a \neq b$. Since $b \neq a', a$ there exists z such that $(a'zb)$ by O4. Hence (zba) and (abz). By Theorem 3, (axb), (abz) imply (axz). Thus we have

$$x \subset a + z, \qquad z \subset a' + b \tag{4}$$

and (1) follows by Definition 3. Now suppose $x = b$. Then (4) holds with the same choice of z and (1) follows as before. Finally suppose $x \subset a' + b$. Then $(a'xb)$ and $x \neq a', a$. Now choose z such that $(a'zx)$. Then (zxa) and (axz). Furthermore $(a'zx)$, $(a'xb)$ imply by Theorem 3 $(a'zb)$. Thus (4) holds and the theorem is established.

THEOREM 8. *If $p \subset a + b$ then $a + b = a + p \cup p \cup b + p$.*

Proof. Suppose $p \subset a + b$. If $a = b$ the result is trivial. Suppose $a \neq b$. Then (apb). Let $R = a + p \cup p \cup b + p$. Suppose $x \subset R$. If $x = p$ then $x \subset a + b$. Suppose $x \subset a + p$. Since $a \neq p$ we have (axp). This with (apb) implies by Theorem 3 (axb), so that $x \subset a + b$. Similarly $x \subset b + p$ implies $x \subset a + b$. Thus $R \subset a + b$. Conversely suppose $x \subset a + b$. Then (axb). If $x = p$ then $x \subset R$. Suppose $x \neq p$. Then (axb), (apb) imply by Theorem 6 (axp) or (apx). If (axp) then $x \subset a + p \subset R$. If (apx), then (axb) implies by Corollary 2 of Theorem 3 (pxb). Thus (bxp) and $x \subset b + p \subset R$. Hence $a + b = R$ and the theorem is proved.

3. The associative law. In this section we show how to extend the definition of $+$ in a spherical geometry S so as to obtain the unrestricted validity of the associative law. However this is impossible within the confines of S (Theorem 9), if S is non-trivial, but can be accomplished very simply in the system formed by the adjunction to S of an "ideal" element o which plays the role of an identity for the operation $+$.

THEOREM 9. *Let spherical geometry S contain at least three points. Then it is impossible to extend our definition of $+$ (Definition 2) to all pairs of points of S in such a way as to preserve the associative law.*[9]

Proof. Suppose such an extension of Definition 2 possible in S—it being understood of course that the iterated sums appearing in the associative law are defined by Definition 3. Suppose $p \neq a, a'$. The associative law and Theorem 7 imply

$$(a + a') + p = a + p \cup p \cup a' + p. \tag{1}$$

Thus $(a + a') + p \supset p$ and by Definition 3 there exists o in S satisfying

$$\text{o} + p \supset p, \qquad a + a' \supset \text{o}. \tag{2}$$

If $p \neq \text{o}, \text{o}'$ the first relation in (2) implies $(\text{o}pp)$. Thus $p = \text{o}$ or o' so that $\text{o} = p$ or p' and (2) implies $a + a' \supset p$ or p'. It is not restrictive to suppose

$$p \subset a + a'. \tag{3}$$

[9] The numerical restriction is essential since the spherical geometry composed of points p, q satisfying $p = q'$, $q = p'$ with vacuous order relation satisfies the associative law if we define $a + a'$ to consist of a, a'.

By O4, $(p'qa)$ for some q. Thus (qap) and $a \subset p + q$. By Definition 3 we may add q to both sides of (3) and we obtain

$$a \subset (a + a') + q. \tag{4}$$

Since $q \neq a,a'$ we may replace p in (1) by q getting in view of (4), $a \subset a + q \cup q \cup a' + q$. This implies (aaq), $a = q$ or $(a'aq)$ which are false, and the proof is complete.[10]

This is not as disappointing as it might seem. The associative law fails in S because it implies (2) which requires that $a + a'$ contain for each p, a "relative" identity element o. This is impossible *in* S, and suggests the possibility of validating (2) and the associative law by going *outside* S. The simplest way to do this is to assign to $a + a'$ an ideal element o, not in S, such that $x + o = o + x = x$ for each x in S. Thus o plays the role of an additive identity and (2) becomes valid. Then if the associative law is to hold

$$a + a' = (a + a) + a' = a + (a + a') \supset a + o = a.$$

Similarly we get $a + a' \supset a'$. Conversely if we require $a + a'$ to consist of a,a',o the associative law holds; we formalize and complete the discussion in the following definition and theorem.

DEFINITION 5. Let S' be the set formed by adjoining to S an "ideal" element o, which is not in S. We extend Definition 2 on sum of *elements* to S' as follows:

$$a + a' = a \cup a' \cup o, \qquad a \subset S;$$
$$b + o = o + b = b, \qquad b \subset S'.$$

Sum of *sets* is determined in S' as in S by Definition 3.[11]

THEOREM 10. *In S' we have* (a) $a+b=b+a$; (b) $(a+b)+c=a+(b+c)$.

Proof. (a) This follows easily from O2, the corollary to Theorem 1 and Definition 5.

(b) The degenerate cases in which one of a,b,c is o or one is the opposite of another can be disposed of using Definition 5, Theorem 7 and O5. Suppose $a,b,c \neq o$ and neither is the opposite of another. Then $a + b$, $b + c$ are significant in the sense of Definition 2, and the result holds by O5 unless one of the following is true:

$$c' \subset a + b, \tag{1}$$
$$a' \subset b + c. \tag{2}$$

But (1) and (2) are equivalent. For (1) implies $a \neq b$ (otherwise $c' = a$) and so $(ac'b)$. Thus $(ac'b) \to (c'ba') \to (ba'c) \to (2)$. Similarly we can show (2) implies (1). Thus we have only to consider the case in which both (1) and (2) hold. In this case we may apply Theorem 8 to $a + b$ and c', and using the associative law for cases already mentioned we have

[10] We have implicitly required that $a + a'$ be non-void since the sum of $a + a'$ and p must be significant by Definition 3, which excludes the void set from consideration. However the theorem is also valid if we allow $a + a'$ to be void.

[11] Note that the converse of Theorem 2 still holds for a,b,c in S with the added proviso $c \neq a'$.

$$
\begin{aligned}
(a+b)+c &= (a+c'\cup c'\cup b+c')+c \\
&= a+c'+c\cup c'+c\cup b+c'+c \\
&= a+(c'\cup c\cup \mathrm{o})\cup c'\cup c\cup \mathrm{o}\cup b+(c'\cup c\cup \mathrm{o}) \\
&= a+c'\cup a+c\cup a\cup c'\cup c\cup \mathrm{o}\cup b+c'\cup b+c\cup b \\
&= a+c'\cup c'\cup b+c'\cup b+c\cup a+c\cup a\cup b\cup c\cup \mathrm{o} \\
&= a+b\cup b+c\cup c+a\cup a\cup b\cup c\cup \mathrm{o}.
\end{aligned}
$$

Since the last expression is symmetrical in a,b,c, when we apply the same argument to $a+(b+c)=(b+c)+a$, as we may in view of (2), we get the same result. Thus (b) is verified.

4. Spherical geometries as multigroups. Continuing the discussion of the last section we show that S', with $+$ as defined, is a multigroup with strong regularity properties and that the theory of spherical geometries is in a sense equivalent to that of a certain class of abelian multigroups.

We begin with the following definition.

DEFINITION 6. A *regular multigroup* is a set G of elements $a,b,c,\ldots$ in which is defined a 2-term operation $+$ satisfying postulates[13] M1, . . . , M5:

M1. *$a+b$ is a uniquely determined non-void subset of G.*

M2. $(a+b)+c=a+(b+c)$.

M3. $a+b=b+a$.

M4. *There exists in G an element* o, *called an identity element, such that $a+\mathrm{o}=a$ for each a in G.*

In G we define $a-b$ to be the set of all x satisfying $b+x\supset a$.

M5. For each b in G there exists b^* in G satisfying[13]

$$a-b=a+b^*. \tag{1}$$

The *order* of regular multigroup G is its cardinal number.

It is easily seen that G has a unique identity element, which may then be represented unambiguously by o. Observe that in view of M5, M1 $a-b\neq O$.[14] M5, M4, M3 imply

$$\mathrm{o}-b=b^*. \tag{2}$$

Thus $\mathrm{o}-b$ is a single element and b^* in M5 is uniquely determined. In view of (2) we naturally call b^* the *negative* or *inverse* of b and denote it $-b$. Thus $-b=\mathrm{o}-b$ and in view of the definition of $\mathrm{o}-b$, we may characterize $-b$ as the unique solution x of the relation $b+x\supset \mathrm{o}$. It easily follows that $-(-b)=b$. Replacing b^* in (1) by $-b$, (1) assumes the form

$$a-b=a+(-b)$$

which is the familiar relation between subtraction and addition of abelian group theory. Thus an abelian group is seen to be a regular multigroup.

[13] We maintain the agreements on identification of elements and unit sets and the use of $\supset$, $\subset$ adopted above (Definition 2) and we extend $+$ from elements to non-void sets by Definition 3.

[13] We are using the term regularity in a much more restricted sense than Dresher and Ore [6, p. 708]; in our sense it implies self-reversibility and complete regularity [6, pp. 717, 723].

[14] O denotes the void set.

We now discuss the relation between spherical geometries and regular multigroups.

THEOREM 11. *If S is a spherical geometry then S', with $+$ as defined, is a regular multigroup in which the negative of point a is its opposite a'.*

Proof. M1,..., M4 hold in S' in view of Definitions 2,5 and Theorem 10. If $b = o$ M5 holds with $b^* = o$ since $a - o = a$. If $b \neq o$ then $b \subset S$ and we choose $b^* = b'$, the *opposite* of b. Then M5 is easily verified if $a = o,b$ or b'. Suppose $a \neq o,b,b'$. Supopse $x \subset a - b$. Then $b + x \supset a$ and $x \neq b,b',o$. Thus $b + x$ is defined by Definition 2 so that by Theorem 2, $b + x \supset a$ implies (bax) and so (axb'). Thus $x \subset a + b'$. Conversely $x \subset a + b' \rightarrow (axb') \rightarrow (b'xa) \rightarrow (xab) \rightarrow (bax) \rightarrow a \subset b + x \rightarrow x \subset a - b$. Thus M5 is completely verified and the theorem is proved.

This result suggests

DEFINITION 7. Let S be a spherical geometry. Then S' with $+$ as defined, is called the *associated multigroup of* S.[15]

The last result does not distinguish associated multigroups of spherical geometries from abelian groups or other regular multigroups. Thus we must find special properties to characterize these multigroups. First we introduce

DEFINITION 8. Let G be a regular multigroup. A *submultigroup* of G is a non-void subset of G which contains with a,b also $-a$ and $a + b$.[16] The *order* of element a of G is the cardinal number of the submultigroup of G generated by a, that is the least submultigroup of G which contains a.

Now we can state and easily derive the characteristic properties of multigroup S'.

THEOREM 12. *The associated multigroup of a spherical geometry is regular, satisfies the idempotent law and each of its elements, with the exception of* o, *has order* 3.

Proof. In view of the last theorem and Definitions 2,5 we have only to show that if S is a spherical geometry and $S' \supset a \neq o$ then the order of a is 3. Any submultigroup of S' which contains a must contain $A = a \cup a' \cup o$, since $-a = a'$ and $a + a' \supset o$. Moreover the negatives of the elements of A are a', $-(a') = a,o$; and A is closed under $+$ in view of the idempotent law and Definition 5. Thus A is the least submultigroup of S' which contains a. The cardinal number of A is 3, since $a,a' \neq o$ and by O3, $a \neq a'$. Thus a has order 3 and the theorem is proved.

Now we prove a sort of converse of this result and characterize the multigroups associated with spherical geometries.

THEOREM 13. *Let G be a regular multigroup which satisfies the idempotent*

[15]Strictly speaking S' is not uniquely determined, since o is not, but we naturally consider the various S' to be identical.

[16]Observe that a submultigroup of G is a regular multigroup with respect to the composition of G. The term submultigroup is often used in a weaker sense than that of Definition 8 to denote a subset which is a multigroup with respect to the composition of the given multigroup [6, p. 714].

law and each element of which, with the exception of o, *has order* 3. *Then G is the associated multigroup of a spherical geometry.*

Proof. Let S be the set obtained by deleting from G its identity element o. In S we define (abc) to mean $c \neq a, -a$ and $b \subset a + c$ and we show that S, with order so defined, is a spherical geometry and that S' its associated multigroup coincides with G.

First we show for $a \neq o$

$$a + (-a) = a \cup (-a) \cup o. \tag{1}$$

Adding a to both members of the relation $o \subset a + (-a)$ we have

$$a \subset a + (a + (-a)) = (a + a) + (-a) = a + (-a).$$

Similarly $-a \subset a + (-a)$; $a, -a, o$ are distinct, for $-a \neq o$ and $a = -a$ implies $a + (-a) = a + a = a$ so that the set $a \cup o$ is the least submultigroup of G containing a, and a has order 2 contrary to hypothesis. Thus since $a + (-a) \supset a, -a, o$ and a has order 3, (1) is verified.

To show S a spherical geometry we observe O1 is a consequence of M5 and (1); O2 follows from M3 and $-(-a) = a$; O3 can be verified by taking p (the opposite of a) to be $-a$, for a in S; O4 follows from M1. To verify O5 consider the operation $\oplus$ defined in S by Definition 2: if $b \neq a, -a$ then $a \oplus b$ is the set of x for which (axb); $a \oplus a = a$. We see immediately that $a \oplus b = a + b$ for $a, b \subset S$ provided $b \neq -a$. Thus since $+$ is associative, the associative law for $\oplus$ certainly holds for those triples a, b, c in S for which it is significant. Hence O5 is verified and S is a spherical geometry.

Now to construct S', the associated multigroup of S, we adjoin o to S to form set S' so that as a *set* $S' = G$. Then we extend $\oplus$ to S' by the agreements (Definition 5) $a \oplus (-a) = a \cup (-a) \cup o$ for $a \subset S$ and $b \oplus o = o \oplus b = b$ for $b \subset S'$. Thus in view of (1) $a \oplus b = a + b$ for all $a, b \subset S'$ and as a *multigroup* $S' = G$.

5. Regular multigroups. In this section we sketch the theory of regular multigroups. The results are analogues of familiar theorems of group theory and are given without proof to avoid duplication of methods in the literature.[17] There is implicit in the discussion, in view of sec. **4**, a corresponding theory for arbitrary spherical geometries, which we explicitly derive for Euclidean spherical geometries. The theory of course also applies to abelian groups. In later sections we add restrictions when necessary and obtain finally the multigroups associated with spherical geometries.

In this section G denotes an arbitrary regular multigroup with elements $a, b, c, \ldots$ and operation $+$; $A, B, C, \ldots$ denote subsets of G which are non-void unless the contrary is stated. For simplicity of expression we shall refer to G as a *group* and to its submultigroups as *subgroups*; and we shall call the usual type of group with *single*-valued composition a *classical* group. The operations

[17] See in particular Dresher and Ore [6]; observe however that many of our definitions differ from theirs.

of subtraction and taking inverses are defined for sets in the natural way: $A - B = \sum_{a \subset A, b \subset B}(a - b)$; $-A$ denotes the set of all $-a$ for $a \subset A$. Familiar formal laws of additive algebra hold for sets: $A \subset A'$, $B \subset B'$ imply $A + B \subset A' + B'$; $(A + B) + C = A + (B + C)$; $A + B = B + A$; $A - B = A + (-B)$; $-(-A) = A$; $-(A + B) = (-A) + (-B)$. Subgroups can be characterized formally. *A is a subgroup of G if and only if* (a) $A + A = A = -A$ *or* (b) $A - A = A$. *Generation* of subgroups is defined in the usual way:

DEFINITION 9. Let M be an arbitrary (not necessarily non-void) subset of G. By the *subgroup of G generated by M*, denoted $\{M\}$, we mean the least subgroup of G which contains M. If $\{M\} = A$ we say M *generates* A or is a *set of generators* of A. In general if M_i, $i \in I$, is a system of arbitrary subsets of G we define $\{M_i; i \in I\}$, the *subgroup of G generated by M_i, $i \in I$*, to be the least subgroup of G which contains each M_i. If I is the set $1, \ldots, n$ we use the notation $\{M_1, \ldots, M_n\}$ for $\{M_i; i \in I\}$. Note that $\{O\} = o$ for any G; if G is the associated multigroup of spherical geometry S then $\{a\} = a \cup (-a) \cup o$, and if S is Euclidean and $a,b \subset S$, $(b \neq a,a')$ then $\{a,b\}$ is the great circle containing a,b to which is adjoined o.

In classical group theory $\{M\}$, where $M \neq O$, consists of all "polynomial" combinations of elements of M which can be formed using the group operation and taking inverses. Here we have an analogous result.

THEOREM 14. *$\{M\}$, if $M \neq O$, is the set union of all expressions $a_1 + \ldots + a_n$ where $a_i \subset M$ or $a_i \subset -M$, $1 \leqq i \leqq n$.*

COROLLARY. (Finiteness of dependence). Suppose $M \neq O$. Then $x \subset \{M\}$ if and only if $x \subset \{a_1, \ldots, a_n\}$ where $a_i \subset M$, $1 \leqq i \leqq n$.

Exactly as in classical abelian group theory we have

THEOREM 15. *If A,B are subgroups of G then $\{A,B\} = A + B$.*

From this Dedekind's famous modular law [4, p. 34, L5] follows, essentially by Dedekind's proof [4, p. 35, Theorem 3.2].

THEOREM 16. (Modularity). *If A,B,C are subgroups of G and $A \subset C$ then*[18] $\{A,B\}.C = \{A,B.C\}$.

Now we point out the geometrical significance of the ideas presented thus far. Let G be the associated multigroup of a Euclidean spherical geometry S and let A be a subgroup of G. By Definition 8, $A \supset a,b$ implies $A \supset -a$, $a + b$. Hence A contains with each point a, its opposite a' and with each pair of points a,b $(b \neq a,a')$ the minor arc of a great circle which joins a and b. An arbitrary (not necessarily non-void) subset of S which enjoys these properties we call a *spherical subspace* or simply a *linear subspace* of S. (Examples are: O, a pair of opposite points, a great circle, etc.) Observe that a linear subspace of S contains with a,b $(b \neq a,a')$ the great circle passing through a,b and so is an analogue in spherical geometry S of a linear subspace of a projective or affine geometry. Let B be the set obtained by deleting o, the

[18] We use the symbol . to denote set theoretic multiplication.

identity element of G, from subgroup A of G. Then B is a linear subspace of S. Furthermore if we adjoin o to B any linear subspace of S we obtain, in view of Definition 5, a corresponding subgroup A of G which we call the subgroup *associated* to B. Thus the trivial operation of adjoining o effects a $(1-1)$ correspondence between the set of linear subspaces of S and the set of subgroups of G, which we call the *natural* correspondence between these sets. In view of this we may consider the concept *linear subspace* of S as essentially identical with *subgroup* of G.[19]

To obtain geometrical significance for $\{M\}$, we suppose $M \not\supset o$, which is not essentially restrictive. If we delete o from $\{M\}$ we obtain a linear subspace $\overline{M}$ of S which, in view of the natural correspondence between linear subspaces of S and subgroups of G, is the *least* linear subspace of S containing M. Thus $\overline{M}$ is called the linear subspace of S *determined* or *spanned* by M. For example the linear subspace of S determined by point a is $a \cup a'$, by $a \cup b$ $(b \neq a,a')$ is the great circle containing a and b. Thus the geometrical notion *determination of linear subspaces* is subsumed under the familiar algebraic concept *generation of subgroups*.[20] Furthermore we note that the natural correspondence associates $\{M\}$ to $\overline{M}$, in particular it associates o to $O, \{a\}$ to the linear space $a \cup a'$, and $\{a,b\}$ to the great circle containing a,b where $b \neq a,a'$.

We continue with *coset* and associated ideas.

DEFINITION 10. Let H be a subgroup of G. Then $a + H$ is called the *coset of H determined by a* and is denoted $(a)_H$. The set of all cosets $(a)_H$ where $a \subset A$ is denoted $(A)_H$. Let G/H denote $(G)_H$. In G/H we define addition thus: $(a)_H \oplus (b)_H = (a + b)_H$. We call G/H with addition so defined the *factor group of G with respect to H*.

As in classical group theory the cosets of H form a decomposition of G. Furthermore the sum of two elements of G/H (cosets) is independent of their representation and G/H, like G, is a *group* (regular multigroup). The correspondence $x \to (x)_H$ maps G on G/H in such a way as to preserve addition. This suggests

DEFINITION 11. Let K_1, K_2 be arbitrary systems (not necessarily groups) consisting of a set of elements and a 2-term operation (not necessarily single-valued) the composition in each being denoted $+$. Let there exist a single-valued mapping f of K_1 on K_2 which satisfies $f(x + y) = f(x) + f(y)$. Then we call f a *homomorphism* of K_1 on K_2 and say K_1 is *homomorphic* to K_2. If f is $(1-1)$ we use the terms *isomorphism*, *isomorphic* and write $K_1 \cong K_2$.[21]

If H is a subgroup of G then G is homomorphic to G/H. Furthermore if G is homomorphic to K, then K also is a group (regular multigroup) and is isomorphic to G/H, where H is the set of elements of G mapped by the homomorphism on the identity of K. If A,B are subgroups of G the mapping

[19] Compare [14, §4], [15, §5].

[20] Compare [14, §5], [15, p. 350, Definition 2].

[21] Congruence relations in groups can be introduced by the definition of [15] and have the familiar relations to homomorphisms [4, pp. 2,3].

$(b)_A \to (b)_{A.B}$ effects an isomorphism of $\{A,B\}/A$ into $B/A.B$ and we may assert

THEOREM 17 (Isomorphism Theorem). *If A,B are subgroups of G then $\{A,B\}/A$ is isomorphic to $B/A.B$.*[22]

From this the Jordan Hölder theorem can be deduced as in classical group theory.

We conclude this section with the geometrical significance of coset and factor group. First let G for simplicity of illustration be the associated multigroup of S, the spherical geometry of a Euclidean 2-sphere, and let H be the subgroup of G formed by adjoining o, the identity element of G, to T a great circle of S. Suppose $a \not\subset H$. Then $(a)_H = a + (o \cup T) = a \cup a + T$. That is $(a)_H$ consists of a and all interior points of minor arcs of great circles which join a to points of T. This is of course the *hemisphere* of S, bounded by T, which contains a. On the other hand $(a)_H = H$ if $a \subset H$. Similarly if we replace T by a pair of opposite points p,p' and let H be the subgroup of G composed of p,p',o we find that the cosets of H are the (open) semicircles with endpoints p,p', and H itself. In general let S be any Euclidean spherical geometry, T be a linear subspace of S, G be the associated multigroup of S and $H = o \cup T$. Then $(a)_H$ is the "hemisphere" bounded by T which contains a provided $a \subset H$, otherwise $(a)_H = H$. Thus the coset concept subsumes the idea *half-space* (point, semicircle, 2-hemisphere, etc.). Furthermore the coset decomposition of G determined by H yields, by exclusion of o from consideration, a decomposition of S into the set of half-spaces (or hemispheres) bounded by T, together with T. Examples are the decomposition of a 2-sphere (1) into a great circle and the hemispheres which it bounds and (2) into a pair of opposite points and the semicircles joining them.

To illustrate the notion factor group consider the second example of the preceding paragraph in which T consists of a pair of opposite points p,p'. Then G/H is the set composed of H and the semicircles joining p and p' in which the "join" or "sum" of two non-opposite semicircles consists of all the semicircles in the *lune* bounded by the given semicircles. G/H is easily seen geometrically to be isomorphic to the multigroup associated with a great circle of S. We prove this formally as a simple application of the Isomorphism Theorem. Let K be the subgroup of G formed by adjoining o to a great circle U which contains neither p nor p'. Then $G = \{H,K\}$ and $H.K = o$ so that by Theorem 17

$$G/H = \{H,K\}/H \cong K/H.K = K/o = K.$$

6. Linear independence and dimension. We continue with the theory of linear independence and dimension or rank which are of importance both in classical group theory and spherical geometry. We consider the assignment of dimension to subgroups of G and its relation to generation and intersection

[22] Compare [6, p. 726, Theorem 6], also see [7, p. 68]. For classical groups see [1, p. 134, Theorem 15], [18, p. 136, the first Isomorphism Theorem].

properties of subgroups. The theory covers the corresponding topics for linear subspaces of a Euclidean spherical geometry and is applicable to spherical geometries in general. The theory requires a restriction on the regular multigroups G we have been studying which relates them in an interesting way to projective geometries (Theorem 19).

In developing the familiar theory of dimension for a Euclidean spherical geometry S we assign to the linear subspaces in order of increasing complexity: O, pair of opposite points, great circle, 2-sphere, . . . , the "dimensions": -1, $0, 1, 2, \ldots$. We may take this to signalize that a linear subspace of S of each type is a maximal proper subspace of one of the succeeding type. Thus a necessary condition for validation of the familiar theory of dimension of S is that O be a maximal proper linear subspace of each pair of opposite points, in other words that there be no linear subspace "between" O and a pair of opposite points a,a'. Translating this into the corresponding restriction on G, the associated multigroup of S, we get since $\{a\}$ is the subgroup of G associated to the linear space composed of a,a': *there is no subgroup of* G *"between"* o *and* $\{a\}$ *if* $a \neq$ o. This property is sufficient to yield the desired dimension theory. In order to phrase it more carefully and conveniently we introduce

DEFINITION 12. Let A, B be distinct subgroups of a regular multigroup G such that $A \supset X \supset B$ (where X is a subgroup of G) implies $X = A$ or $X = B$. Then we say A *covers* B.

We state the desired property of a regular multigroup G which we assume throughout this section as the

COVERING POSTULATE. If $a \neq$ o, $\{a\}$ covers o.

We continue with consequences of this postulate, postponing to the end of the section interpretations of the theory. Suppose $\{a\} \supset b \neq$ o. Then $\{a\} \supset \{b\} \supset$ o, and $\{b\} \neq$ o. Hence by the Covering Postulate $\{b\} = \{a\}$. Thus we may assert the

COROLLARY. If $a \neq$ o, $\{a\}$ is generated by each of its elements other than o.

We generalize the Covering Postulate in

THEOREM 18. *If* H *is a subgroup of* G *and* $a \not\subset H$ *then* $\{a,H\}$ *covers* H.

Proof. Suppose $H \subset X \subset \{a,H\}$ where $X \neq H$ and is a subgroup of G. Suppose $x \subset X$, $x \not\subset H$. Then $x \subset \{\{a\},H\} = \{a\} + H$ so that $x \subset b + h$ where $b \subset \{a\}$ and $h \subset H$. If $b =$ o then $x = h$ contrary to $x \not\subset H$. Thus $b \neq$ o. We have $b \subset x - h \subset X$. Thus using the last corollary $X \supset \{b\} = \{a\}$. Hence $X \supset \{a,H\}$ and $X = \{a,H\}$. Since $\{a,H\} \neq H$, by definition $\{a,H\}$ covers H.

It is well known that a Euclidean sphere is convertible into a projective geometry by defining "point" as a pair of opposite points of the sphere, and "line" as the set of "points" contained in a great circle. The following theorem which we shall not prove is, in essence, a generalization of this and implies that spherical geometries are related to projective geometries in essentially the same way.

THEOREM 19. *Let P be the set of subgroups of G of the form $\{a\}$, $a \neq o$. Then P becomes a projective geometry if we define "point" to be element of P and "line" to be the set of "points" contained in a subgroup of G of the form $\{a,b\}$, where $\{a\} \neq \{b\}$.*[33]

The Covering Postulate implies that G has marked homogeneity of structure.

THEOREM 20. *If $a,b \neq o$ then $\{a\} \cong \{b\}$.*

Proof. Suppose $\{a\} \neq \{b\}$, since otherwise the result is trivial. Suppose $o \neq c \subset a + b$. Then $\{c\} \neq \{a\}$ for otherwise $\{a\} \supset c - a \supset b$ and by the corollary to the Covering Postulate $\{a\} = \{b\}$. We have $\{a\} \subset \{a,c\} \subset \{a,b\}$ so that $\{a,c\} = \{a,b\}$ by Theorem 18. By symmetry $\{b,c\} = \{a,b\}$ so that $\{a,c\} = \{b,c\}$. We have, using the Isomorphism Theorem

$$\{a,c\}/\{c\} = \{\{a\},\{c\}\}/\{c\} \cong \{a\}/(\{a\}.\{c\}) = \{a\}/o = \{a\}.$$

By symmetry $\{b,c\}/\{c\} \cong \{b\}$. Thus $\{a,c\} = \{b,c\}$ implies $\{a\} \cong \{b\}$.

COROLLARY. If A covers B, A' covers B' then $A/B \cong A'/B'$.

In choosing a set of generators for a group we naturally want to exclude redundant elements. This suggests the following definition of *linear independence.*

DEFINITION 13. M an arbitrary (not necessarily non-void) subset of G is *linearly independent* or *independent* if[34] $\{M \dot{-} x\} \not\supset x$ for each $x \subset M$. In the contrary case we say M is *dependent.* Observe that O is independent but o is dependent.

If M is dependent then the corollary to Theorem 14 implies that some finite subset of M is likewise dependent. The converse is obvious. Thus we may state

THEOREM 21. *M is independent if and only if its finite subsets are independent.*[35]

We now derive a criterion for independence of a finite set very similar to the familiar algebraic one for independence of elements of a linear vector space or an abelian group.

THEOREM 22. *Suppose $a_1, \ldots, a_n$ are distinct and $a_i \neq o$, $1 \leqq i \leqq n$. Then they constitute an independent set if and only if*

$$p_1 + \ldots + p_n \supset o, \qquad p_i \subset \{a_i\}, \qquad (1 \leqq i \leqq n)$$

always implies $p_1 = \ldots = p_n = o$.

Proof. Suppose $a_1, \ldots, a_n$ distinct and form an independent set, $p_i \subset \{a_i\}$, $1 \leqq i \leqq n$, $p_1 + \ldots + p_n \supset o$, but one of the p's, say $p_n \neq o$. Then $p_1 + \ldots + p_{n-1} \supset -p_n \neq o$ and using the corollary to the Covering Postulate we have

$$\{a_1, \ldots, a_{n-1}\} \supset \{-p_n\} = \{a_n\} \supset a_n,$$

[33] Compare Carmichael [5, Chap. XI] where finite projective geometries are represented by systems of subgroups of a certain type of finite abelian group.

[34] We use the symbol $\dot{-}$ to denote set theoretic subtraction.

[35] Compare [12, Theorem 2], [13, Theorem 2.3].

contrary to supposition. Conversely suppose $a_1, \ldots, a_n$ satisfy the given condition, each distinct from o, but they do not form an independent set. It is not restrictive to assume $\{a_1, \ldots, a_{n-1}\} \supset a_n$. Then by Theorem 15, $a_n \subset \{a_1\} + \ldots + \{a_{n-1}\}$ so that there exist $p_i \subset \{a_i\}$, $1 \leqq i \leqq n-1$, satisfying $a_n \subset p_1 + \ldots + p_{n-1}$. Adding $-a_n$ to both members of this relation we obtain $o \subset p_1 + \ldots + p_n$, where $p_n = -a_n \neq o$, contrary to supposition and the proof is complete.

In view of the corollary to the Covering Postulate, if $M \not\supset o$ it is independent if and only if $\{M \dot{-} x\} . \{x\} = o$ for $x \subset M$. This property is now strengthened.

THEOREM 23. *Suppose $M \not\supset o$. Then M is independent if and only if M_1, $M_2 \subset M$ and $M_1.M_2 = O$ always imply $\{M_1\}.\{M_2\} = o$.*[26]

Proof. Supose M independent, M_1, $M_2 \subset M$ and $M_1.M_2 = O$ but $\{M_1\}.\{M_2\} \neq o$. Then $o \neq p \subset \{M_1\}.\{M_2\}$ for some p. Thus $M_1 \neq O$ and by the corollary to Theorem 14 there exist $a_i \subset M_1$, $1 \leqq i \leqq n$, such that $p \subset \{a_1, \ldots, a_n\}$. We may assume that in this relation redundant a's have been deleted. Thus a_n, $p \not\subset \{a_1, \ldots, a_{n-1}\}$.[27] Hence by Theorem 18,

$$\{a_1, \ldots, a_n\} \text{ covers } \{a_1, \ldots, a_{n-1}\} \text{ and}^{28}$$

$$a_n \subset \{a_1, \ldots, a_n\} = \{a_1, \ldots, a_{n-1}, p\} \subset \{a_1, \ldots, a_{n-1}, M_2\} \subset \{M \dot{-} a_n\},$$

contrary to supposition and the necessity of the condition is proved. Its sufficiency is immediate if $M \not\supset o$, since it implies $\{M \dot{-} x\}.\{x\} = o$ for $x \subset M$.

The theory of dimension for subgroups of G is covered by the theory of exchange lattices of MacLane [12] since Theorems 16, 18 imply the subgroups of G form a modular exchange lattice. We have the following results. Each subgroup A of G has a *basis*, that is an independent set of generators. Any two bases of A have the same cardinal number, which we call the *dimension* or *rank* of subgroup A, denoted functionally $d(A)$.[29] If B also is a subgroup of G, $A \supset B$ implies $d(A) \geqq d(B)$. For each subgroup A there exists a *complement*, A', that is a subgroup of G such that $\{A,A'\} = G$, $A.A' = o$. For subgroups A,B of finite dimension we have the *dimension formula*

$$d(\{A,B\}) + d(A.B) = d(A) + d(B).$$

Furthermore if $A \supset B$, the relation A covers B is equivalent to $d(A) = d(B) + 1$. For finite n, a set of n independent elements is contained in a unique subgroup of G of dimension n. Finally, if $d(A) = n$ is finite, any independent set of n elements of A is a basis of A.

Now we consider applications of the theory developed in this section. It certainly applies to the associated multigroup of a spherical geometry since in

[26] Compare [13, Definition 2.1].

[27] If $n = 1$ this expression stands for $\{a_i; i \epsilon O\} = o$.

[28] If $n = 1$ we naturally take the expression $\{a_1, \ldots, a_{n-1}, p\}$ to be $\{p\}$.

[29] In applying MacLane's theory [12, Theorem 6] $d(A)$ would be defined as the cardinal number of a basis of A in the lattice of subgroups of G. Using Theorem 23 this can be shown equivalent to our definition of $d(A)$.

this case $\{a\}$ consists of $a, -a, o$ and the Covering Postulate obviously applies. Thus the results of the last paragraph yield a theory of linear independence and dimension for spherical geometries in general and, in view of the discussion in sec. 5 of the geometrical significance of subgroups, cover the theory of alignment and intersection of linear subspaces of a Euclidean spherical geometry.[30]

Next we naturally enquire which classical abelian groups G satisfy the theory, that is to which does the Covering Postulate apply. Suppose then that G is a classical abelian group satisfying the Covering Postulate. Clearly the cyclic subgroups $\{a\}$ of G, where $a \neq o$, must have prime order, and by Theorem 20 all must have the same order p. The existence of a basis M of G implies that G is a direct sum of cyclic groups of order p. In fact G is the direct sum [2] of the system of cyclic subgroups $\{x\}$, $x \subset M$. For in the first place $G = \{M\} = \{\{x\}; x \subset M\}$, that is G is generated by this system of groups. Secondly the intersection of each group of the system with the group generated by the remaining groups of the system is o, since $\{a\}.\{\{x\}; x \subset M \dot{-} a\} = \{a\}.\{M \dot{-} a\} = o$ by Theorem 23. Conversely any direct sum of (classical) cyclic groups of prime order p satisfies the Covering Postulate since each of its cyclic subgroups, other than the identity, has order p. Thus the theory of multigroups developed so far relates Euclidean spheres and direct sums of (classical) cyclic groups of fixed prime order, which agree in the more general group theoretic properties of the preceding section as well as in the dimensional properties of this section.[31]

7. Separation and factor groups. In this final section we derive conditions that the familiar type of separation theory, which holds for the linear spaces of a Euclidean spherical geometry, be valid in a regular multigroup, and show in effect (Theorem 24, Corollary 4) that this theory holds for any spherical geometry.

In this section G will denote an arbitrary regular multigroup, all other restrictions on G will be stated explicitly. We begin with the precise sense in which the term *separation* will be used in G.

DEFINITION 14. Let A,B be subgroups of G and let X,Y exist such that (1) $A = B \cup X \cup Y$, $B.X = X.Y = B.Y = O$, $X,Y \neq O$ and (2) $x_1,x_2 \subset X$, $y_1,y_2 \subset Y$ imply $x_1 + x_2 \subset X$, $y_1 + y_2 \subset Y$, $(x_1 + y_1).B \neq O$. Then we say B *separates* A.

The theory of separation based on this definition is independent of the dimension of the subgroups involved, which may be finite or infinite. We begin with a basic criterion for separation in terms of factor group.

[30] If G is the associated multigroup of a Euclidean spherical geometry S and A is the subgroup of G associated to linear subspace T of S, then $d(A)$ exceeds by unity the dimension of T as ordinarily defined.

[31] This is related in view of Theorem 19 to results of Carmichael [5] on the representation of finite projective geometries by systems of subgroups of finite groups of the type mentioned. For deep analogies between projective geometries and classical abelian groups see Baer [3].

THEOREM 24. *B separates A if and only if A/B is isomorphic to the group G_3 of order 3 whose addition table is the following:*[32]

	o	p	$-p$
o	o	p	$-p$
p	p	p	o,p,$-p$
$-p$	$-p$	o,p,$-p$	$-p$

Proof. Suppose B separates A, and that X,Y satisfy the conditions of Definition 14. We show

$$A = B \cup X \cup Y \tag{1}$$

is the coset decomposition of A determined by its subgroup B. First we suppose $a \subset X$ and show $X = a + B$. We have $-a \subset Y$. For $-a \subset X$ implies $o \subset a + (-a) \subset X$ contrary to $o \subset B$; and $-a \subset B$ implies $a \subset B$ contrary to $a \subset X$. Hence for arbitrary $x \subset X$ we have $x + (-a) \supset b$ for some $b \subset B$. Thus $x \subset a + b \subset a + B$. Conversely for arbitrary $x \subset a + B$ we have $x \subset a + b$ for some $b \subset B$. If $x \subset Y$ then $b \subset x - a = x + (-a) \subset Y$ contrary to $b \subset B$; if $x \subset B$ then $a \subset x - b \subset B$ contrary to $a \subset X$. Hence $x \subset X$ and $X = a + B$. By symmetry since $-a \subset Y$ we have $Y = (-a) + B$. Thus (1) becomes

$$A = B \cup a + B \cup (-a) + B,$$

and A/B is composed of the cosets $(a)_B$, $(-a)_B$, $(o)_B$. We determine the addition table of A/B. We have $(a)_B \oplus (a)_B = (a + a)_B$. Since $a + a \subset X = (a)_B$, we have $(a + a)_B = (a)_B$ so that $(a)_B \oplus (a)_B = (a)_B$. Likewise $(-a)_B \oplus (-a)_B = (-a)_B$. Since $(a)_B \oplus (-a)_B \supset (o)_B$, we have adding $(a)_B$ to both members, $(a)_B \oplus (-a)_B \supset (a)_B$. Similarly $(a)_B \oplus (-a)_B \supset (-a)_B$. Thus since $(a)_B$, $(-a)_B$, $(o)_B$ are distinct, A/B is easily seen to be isomorphic to G_3.

Conversely suppose A/B isomorphic to G_3. Let $A = B \cup X \cup Y$ be the coset decomposition of A determined by B. Then $B \cdot X = X \cdot Y = B \cdot Y = O$ and $X, Y \neq O$. Since $A/B \cong G_3$ we have in A/B, $X \oplus X = X$, $Y \oplus Y = Y$, $X \oplus Y \supset B$. Thus if $x_1, x_2 \subset X$ we have $X \oplus X = (x_1)_B \oplus (x_2)_B = (x_1 + x_2)_B = X$, so that $x_1 + x_2 \subset X$. Similarly $y_1, y_2 \subset Y$ imply $y_1 + y_2 \subset Y$. Finally $X \oplus Y = (x_1)_B \oplus (y_1)_B = (x_1 + y_1)_B \supset B$ so that $x_1 + y_1 \supset b$ for some $b \subset B$. Thus B separates A by definition.

COROLLARY 1. *B separates A* implies *A covers B.*

Proof. The hypothesis implies $A/B \cong G_3$. Since G_3 covers its identity, A/B has the same property and the conclusion follows easily.[33]

The typical separation property of linear subspaces of a Euclidean spherical geometry (or of a Euclidean *space* for that matter) suggests the converse of the corollary, namely: *A covers B implies B separates A.*[34] We seek conditions

[32] Observe that G_3 is $\{p\}$ if $p \neq o$ is an element of the associated multigroup of any spherical geometry.

[33] For classical groups compare [1, p. 134, Theorem 14].

[34] A similar property holds in descriptive geometries [15, p. 372, Theorem 6].

that this hold in G. In view of Theorem 24, the desired property is equivalent to: *A covers B implies $A/B \cong G_3$*. Suppose G satisfies the Covering Postulate and to exclude trivial cases suppose $G \neq o$. Then by the corollary to Theorem 20 all factor groups A/B, where A covers B, are isomorphic; thus all are isomorphic to G_3 if (and only if) one is. This one may be chosen arbitrarily. Taking it to be $\{a\}/o = \{a\}$, where $a \neq o$, we have

COROLLARY 2. Let $G \neq o$ satisfy the Covering Postulate. Then A covers B implies B separates A, and if and only if G has a subgroup isomorphic to G_3.

Suppose G satisfies the Covering Postulate and has a subgroup isomorphic to G_3. This subgroup must be of the form $\{a\}$, $a \neq o$, so that by Theorem 20 all subgroups of this form are isomorphic to G_3. But the latter condition implies the Covering Postulate. Thus we may reformulate the sufficiency in Corollary 2 as

COROLLARY 3. Suppose all subgroups of G of the form $\{a\}$, $a \neq o$, are isomorphic to G_3. Then A covers B implies B separates A.

Finally we observe that $\{a\}$ is isomorphic to G_3 if and only if a has order 3 and $a + a = a$ (see the derivation of (1) in the proof of Theorem 13). Thus we have

COROLLARY 4. Suppose G satisfies the idempotent law and each of its elements, with the exception of o, has order 3. Then A covers B implies B separates A.

In view of Theorem 12 this result yields a separation theory for spherical geometries.

REFERENCES

[1] A. A. Albert, *Modern Higher Algebra* (Univ. Chicago press, 1937).
[2] R. Baer, *Lectures on Abelian Groups*, mimeographed. Institute for Advanced Study, 1936.
[3] ———, "A Unified Theory of Projective Spaces and Finite Abelian Groups," *Trans., Amer. Math. Soc.*, vol. 52 (1942), 283-343.
[4] G. Birkhoff, *Lattice Theory*. Amer. Math. Soc. Colloquium pub., 1940.
[5] R. D. Carmichael, *Theory of Groups of Finite Order* (Ginn, 1937).
[6] M. Dresher and O. Ore, "Theory of Multigroups," *Amer. J. Math.*, vol. 60 (1938), 705-733.
[7] J. E. Eaton and O. Ore, "Remarks on Multigroups," *Amer. J. Math.*, vol. 62 (1940), 67-71.
[8] D. A. Flanders, "Double Elliptic Geometry in Terms of Point, Order and Congruence," *Ann. of Math.*, vol. 28 (1926-27), 534-548.
[9] G. H. Hallett, Jr., "Linear Order in Three Dimensional Euclidean and Double Elliptic Spaces," *Ann. of Math.*, vol. 21 (1921), 185-202.
[10] G. B. Halsted, *Rational Geometry* (Wiley, 1904).
[11] J. R. Kline, "Double Elliptic Geometry in Terms of Point and Order Alone," *Ann. of Math.*, vol. 18 (1916-17), 31-44.
[12] S. MacLane, "A Lattice Formulation for Transcendence Degrees and p-bases," *Duke Math. J.*, vol. 4 (1938), 455-468.
[13] J. von Neumann, *Continuous Geometry*, part I, mimeographed. Institute for Advanced Study, 1936.
[14] W. Prenowitz, "Projective Geometries as Multigroups," *Amer. J. Math.*, vol. 65 (1943), 235-256.

[15] ———, "Descriptive Geometries as Multigroups," *Trans. Amer. Math. Soc.*, vol. 59 (1946), 333-380.
[16]———, "Total Lattices of Convex Sets and of Linear Spaces," *Ann. of Math.*, vol. 49 (1948), 659-688.
[17] O. Veblen, *The Foundations of Geometry*, in *Monographs on Topics of Modern Mathematics*, edited by J. W. A. Young (Longmans, 1915).
[18] B. L. van der Waerden, *Moderne Algebra*, vol. 1, 1st ed. (Springer, 1930).

Brooklyn College

THE BIANCHI IDENTITIES IN THE GENERALIZED THEORY OF GRAVITATION

A. EINSTEIN

1. General remarks. The heuristic strength of the general principle of relativity lies in the fact that it considerably reduces the number of imaginable sets of field equations; the field equations must be covariant with respect to all continuous transformations of the four coordinates. But the problem becomes mathematically well-defined only if we have postulated the dependent variables which are to occur in the equations, and their transformation properties (field-structure). But even if we have chosen the field-structure (in such a way that there exist sufficiently strong relativistic field-equations), the principle of relativity does not determine the field-equations uniquely. The principle of "logical simplicity" must be added (which, however, cannot be formulated in a non-arbitrary way). Only then do we have a definite theory whose physical validity can be tested *a posteriori*.

The relativistic theory of gravitation bases its field-structure on a symmetric tensor g_{ik}. The most important physical reason for this is that in the special theory we are convinced of the existence of a "light-cone" ($g_{ik}dx^i dx^k = 0$) at each world-point, which separates space-like line-elements from time-like ones. What is the most natural way of generalizing this field-structure? The use of a non-symmetric tensor seems to be the simplest possibility, although this cannot be justified convincingly from a physical standpoint. But the following formal reason seems to me important. For the general theory of gravitation it is essential that we can associate with the covariant tensor g_{ik} a contravariant g^{ik}, through the relation $g_{i\alpha}g^{k\alpha} = \delta_i^{\,k} = g_{\alpha i}g^{\alpha k}$ (normalized cofactors). This association can be carried over to the non-symmetric case directly. So it is natural to try to extend the theory of gravitation to non-symmetric g_{ik}-fields.

The main difficulty in this attempt lies in the fact that we can build many more covariant equations from a non-symmetric tensor than from a symmetric one. This is due to the fact that the symmetric part, $g_{(ik)}$, and the antisymmetric part, $g_{[ik]}$, are tensors independently. Is there a formal point of view which makes one of the many possibilities seem most natural? It seems to me that there is. In the case of the gravitational theory it is essential that besides the g_{ik} tensor we also have the symmetric infinitesimal displacement $\Gamma_{ik}^{\,l}$. This is connected with g_{ik} by the equation

$$g_{ik,l} - g_{\alpha k}\Gamma_{il}^{\,\alpha} - g_{i\alpha}\Gamma_{lk}^{\,\alpha} = 0. \tag{1}$$

But in the symmetric case the order of indices does not matter. How shall we generalize (1) to our case? We make use of the following postulate: there

Received March 12, 1949.

is a tensor g_{ki}, the "conjugate" of g_{ik}, and a Γ_{ki}^{l} "conjugate" of Γ_{ik}^{l}. It seems reasonable that conjugates should play equivalent roles in the field-equations. So we require that if in any field-equation we replace g and Γ by their conjugates, we should get an equivalent equation. This requirement replaces symmetry in our system. (See sec. 2.) If we require that the set of equations (1) should go over into itself under this operation of "conjugation," then the order of indices must be as in (1).

Our main task now is to find out whether there is a sufficiently convincing method of finding a unique set of field-equations for the non-symmetric fields with the above structure. In both previous publications[1] this was solved by forming a variational principle in close analogy to the symmetric case. This way we make sure that the resulting equations will be compatible. The only reason why this derivation may seem not completely satisfactory is that we subject the field *a priori* to two conditions, for reasons of logical simplicity:

$$\Gamma_{\underset{\smile}{ik}}{}^{s} = \tfrac{1}{2}(\Gamma_{ik}{}^{s} - \Gamma_{si}{}^{s}) = 0, \tag{2}$$

$$\mathfrak{g}^{\underset{\smile}{ik}}{}_{,s} = \tfrac{1}{2}(\mathfrak{g}^{ik} - \mathfrak{g}^{si})_{,s} = 0; \quad (\mathfrak{g}^{ik} = g^{ik}(-\det g_{ab})^{\frac{1}{2}}). \tag{3}$$

These side-conditions make the derivation more complex than in the gravitational theory, and their formal justification has not been accomplished in a fully satisfactory manner so far.[2]

In the theory of symmetric fields there is a second method of ensuring the compatibility of the field-equations ($R_{ik} = 0$). We must have four identities connecting the equations. These can be derived by contracting the Bianchi-identities which hold for the curvature tensor:

$$R_{\underset{\cdot\cdot}{ik}lm;n} = R_{ikl\underset{\cdot}{m};n} + R_{ikm\underset{\cdot}{n};l} + R_{ikn\underset{\cdot}{l};m} = 0.$$

In this article we shall show that an analogous argument can be used for the justification of the field-equations also in our case. This will give a deeper insight into the structure of non-symmetric fields, and it will demonstrate in a new way that the field-equations chosen for the non-symmetric fields are really the natural ones.

2. Non-symmetric tensors. For the sake of convenient reference we shall sum up the main facts of the calculus of non-symmetric tensors.

Given any tensor A_{ik}, it can be written as the sum of a symmetric tensor $A_{\underline{ik}}$ and an antisymmetric $A_{\underset{\smile}{ik}}$. These are uniquely determined by the relations:

$$A_{\underline{ik}} = \tfrac{1}{2}(A_{ik} + A_{ki}), \tag{4}$$

$$A_{\underset{\smile}{ik}} = \tfrac{1}{2}(A_{ik} - A_{ki}). \tag{5}$$

A complication is introduced into this theory by the fact that besides the fundamental tensor g_{ik} we also have its conjugate

$$\bar{g}_{ik} = g_{ki}. \tag{6}$$

[1] *Ann. of Math.*, vol. 46 (1945), no. 4; vol. 47(1946), no. 4.

[2] It is a consequence of (1) that (2) and (3) are equivalent. This will be proven in sec. 5.

The other tensors of our theory are defined in terms of g_{ik}. Given a tensor A_{ik}, by its *conjugate* $\bar{A}_{ik}$ we mean[3] the tensor we get by replacing g_{ik} in the definition of A_{ik} by g_{ki}. (This definition agrees with (6) in particular.) We shall be particularly interested in tensors in whose definition g and $\bar{g}$ play analogous roles; more precisely those tensors for which replacing g_{ik} by g_{ki} merely changes A_{ik} into A_{ki}, or for which

$$\bar{A}_{ik} = A_{ki}. \tag{7}$$

A tensor having the property (7) is called *Hermitian*.[3] More generally any function $A \ldots_{ik} \ldots$ of the g_{ik} is Hermitian in (ik) if

$$\bar{A} \ldots_{ik} \ldots = A \ldots_{ki} \ldots \tag{7a}$$

If $\Gamma_{ik}{}^{l}$ is defined by (1), then Γ is Hermitian in (ik). This is another way of stating the principle by which we chose the order of indices in (1).[4]

We say that $A \ldots_{ik} \ldots$ is anti-Hermitian if

$$\bar{A} \ldots_{ik} \ldots = -A \ldots_{ki} \ldots \tag{8}$$

In analogy to (4), (5), we can decompose any tensor uniquely into

$$A_{ik} = \tfrac{1}{2}(A_{ik} + \bar{A}_{ki}) + \tfrac{1}{2}(A_{ik} - \bar{A}_{ki}). \tag{9}$$

The first term is the Hermitian, the second the anti-Hermitian part of A_{ik}.

Covariant derivatives still have to be generalized. In the symmetric theory, if $A \ldots^{i}_{k} \ldots$ is any tensor, then

$$A \ldots^{i}_{k} \ldots {}_{;l} = A \ldots^{i}_{k} \ldots {}_{,l} \pm \cdots + A \ldots^{s}_{k} \ldots \Gamma_{sl}{}^{i} - A \ldots^{i}_{s} \ldots \Gamma_{kl}{}^{s} \pm \cdots$$

is also a tensor. This is true in our theory also, but we can order the two lower indices of Γ in two ways, in each term (after the first one). If the differentiation index l is to be on the right in a certain term, we put $+$ under the corresponding tensor-index; if on the left, put $-$ under the index. As an illustration we give a new form of (1):

$$g_{\underset{+}{i}\,\underset{-}{k};l} = g_{ik,l} - g_{sk}\Gamma_{il}{}^{s} - g_{is}\Gamma_{lk}{}^{s} = 0. \tag{1a}$$

The theorems about covariant differentiation can be taken over from the symmetric theory, if we are careful to distinguish the two kinds of derivatives. By raising the indices i and k in (1a) we have:

$$g^{\underset{-}{i}\,\underset{+}{k}}{}_{;l} = g^{ik}{}_{,l} + g^{sk}\Gamma_{sl}{}^{i} + g^{is}\Gamma_{ls}{}^{k} = 0. \tag{1b}$$

Sometimes it is even convenient to write things like

$$A_{\underset{-}{i}\,\underset{+}{k}\,lm;n} = A_{iklm,n} - A_{sklm}\Gamma_{ni}{}^{s} - A_{islm}\Gamma_{kn}{}^{s}$$

---

[3] The names "conjugate" and "Hermitian" can be justified as follows: an interesting possibility is to choose g_{ik} imaginary. Then $\bar{g}$ is really the conjugate of g. Hence A is the conjugate of A, and the definition of "Hermitian" agrees with the usual one.

[4] Thus in our theory the condition of symmetry is generalized to that of being Hermitian. g_{ik}, $\Gamma_{ik}{}^{l}$, R_{ik} are all Hermitian in (ik).

but it must be remembered that such expressions are not always tensors, unless + or − occurs under *each* tensor subscript.

If we let $\mathfrak{g}$ stand for the square-root of the negative determinant of g_{ik}, then $\mathfrak{g}$ is a scalar density. We can describe a tensor density as a product of $\mathfrak{g}$ and a tensor. Let us study these densities. Multiply (1) by g^{ik} and sum:

$$\frac{(\det g_{ik}),_l}{(\det g_{ik})} - \Gamma_{sl}{}^{s} - \Gamma_{ls}{}^{s} = 0,$$

$$\frac{(\mathfrak{g}^2),_l}{\mathfrak{g}^2} - 2\Gamma_{l\underline{s}}{}^{s} = 0,$$

$$\mathfrak{g},_l - \mathfrak{g}\Gamma_{l\underline{s}}{}^{s} = 0. \tag{10}$$

It is, therefore, natural to define[5] $\mathfrak{g}_{;l}$ as $\mathfrak{g},_l - \mathfrak{g}\Gamma_{l\underline{s}}{}^{s}$.

If (1a) is satisfied, then $\mathfrak{g}_{;l} = 0$. If we do not assume (1), then $g_{\underset{+-}{i\,k};l}$ and $g^{\underset{+-}{i\,k}}{}_{;l}$ do not vanish but they have tensorial character. Also $\mathfrak{g}_{;l}$ has then the character of a vector density.

We can now calculate the covariant derivative of a tensor density from the rule for differentiating a product. For example:

$$\mathfrak{g}^{ik}{}_{;l} = (\mathfrak{g}g^{ik})_{;l} = \mathfrak{g}_{;l}\, g^{ik} + \mathfrak{g}g^{ik}{}_{;l}.$$

This vanishes, if (1) is satisfied. More explicitly:

$$\begin{aligned}\mathfrak{g}^{ik}{}_{;l} &= (\mathfrak{g},_l - \mathfrak{g}\Gamma_{l\underline{s}}{}^{s})\, g^{ik} + \mathfrak{g}(g^{ik},_l + g^{sk}\Gamma_{sl}{}^{i} + g^{is}\Gamma_{ls}{}^{k}) \\ &= \mathfrak{g}^{ik},_l + \mathfrak{g}^{sk}\Gamma_{sl}{}^{i} + \mathfrak{g}^{is}\Gamma_{ls}{}^{k} - \mathfrak{g}^{ik}\Gamma_{l\underline{s}}{}^{s}.\end{aligned}$$

Therefore we have:

$$\mathfrak{g}^{\underset{+-}{i\,k}}{}_{;l} = \mathfrak{g}g^{\underset{+-}{i\,k}}{}_{;l} = 0.$$

For completeness we include the following abbreviation:

$$A_{\underset{\cdots}{ikl}} = A_{ikl} + A_{kli} + A_{lik}.$$

3. Properties of the generalized curvature. We start with a non-symmetric Γ and build the curvature tensor as usual by parallel translation of a vector around an infinitesimal area element:

$$R^{i}{}_{klm} = \Gamma_{kl}{}^{i},_m - \Gamma_{km}{}^{i},_l - \Gamma_{sl}{}^{i}\Gamma_{km}{}^{s} + \Gamma_{sm}{}^{i}\Gamma_{kl}{}^{s}. \tag{11}$$

A direct computation shows that the tensor satisfies the identities:

$$R^{\underset{+}{i}}{}_{\underset{\cdots}{klm};n} = R^{i}{}_{\underset{\cdots}{klm},n} + R^{s}{}_{\underset{\cdots}{klm}}\Gamma_{\underset{\cdot}{sn}}{}^{i} - R^{i}{}_{\underset{\cdots}{slm}}\Gamma_{\underset{\cdot}{kn}}{}^{s} = 0. \tag{12}$$

From (11) we can form the covariant curvature tensor in analogy to the symmetric case,

$$R_{iklm} = g_{si}R^{s}{}_{klm}. \tag{13}$$

The choice of g_{si} instead of g_{is} may seem arbitrary, but this is not really so. We have to lower the index i in the identities (12). The contravariant index i has the + differentiation character, so it must be summed with a similar

[5] Since $\Gamma_{l\underline{s}}{}^{s} = \Gamma_{sl}{}^{s}$, the two kinds of differentiation coincide when applied to $\mathfrak{g}$. This must be so since there is no index which could have a + or − character.

index, i.e. the first index of g. Only this way can we lower the index i in (12) without introducing additional terms. Thus we get the covariant identities

$$g_{si}R^{i}_{\underset{+}{k}\underset{\cdot}{l}\underset{\cdot}{m};\underset{\cdot}{n}} = (g_{si}R^{s}_{\underset{-}{k}\underset{+}{l}\underset{\cdot}{m}})_{;\underset{\cdot}{n}} = R_{\underset{-}{i}\underset{+}{k}\underset{\cdot}{l}\underset{\cdot}{m};\underset{\cdot}{n}} = 0. \tag{14}$$

For what follows we must also find the symmetry properties of R_{iklm}. From (11) it is clear that R^{i}_{klm} is antisymmetric in (lm). From (13) we see that R_{iklm} has the same property:

$$R_{iklm} = -R_{ikml}. \tag{15}$$

If we differentiate (1) with respect to m and antisymmetrize with respect to l and m, we have

$$(g_{ik,l} - g_{sk}\Gamma_{il}^{s} - g_{is}\Gamma_{lk}^{s})_{,m} - (g_{ik,m} - g_{sk}\Gamma_{im}^{s} - g_{is}\Gamma_{mk}^{s})_{,l} = 0$$

or

$$\begin{aligned} &- g_{sk,m}\Gamma_{il}^{s} - g_{is,m}\Gamma_{lk}^{s} + g_{sk,l}\Gamma_{im}^{s} + g_{is,l}\Gamma_{mk}^{s} \\ &- g_{sk}(\Gamma_{il}^{s}{}_{,m} - \Gamma_{im}^{s}{}_{,l}) - g_{is}(\Gamma_{lk}^{s}{}_{,m} - \Gamma_{mk}^{s}{}_{,l}) = 0. \end{aligned}$$

Using (1) again on the first four terms and then collecting terms,

$$\begin{aligned} &-g_{sk}(\Gamma_{il}^{s}{}_{,m} - \Gamma_{im}^{s}{}_{,l} - \Gamma_{il}^{s}\Gamma_{im}^{t} + \Gamma_{tm}^{s}\Gamma_{il}^{t}) \\ &\qquad - g_{is}(\Gamma_{lk}^{s}{}_{,m} - \Gamma_{mk}^{s}{}_{,l} - \Gamma_{ll}^{s}\Gamma_{mk}^{t} + \Gamma_{mt}^{s}\Gamma_{lk}^{t}) = 0 \end{aligned}$$

or, using (11), (13), we have

$$R_{kilm} = -R_{iklm}. \tag{16}$$

This expresses that R_{iklm} is anti-Hermitian in (ik); this is the manner in which the antisymmetry of R_{iklm} (in the gravitational theory) generalizes to our case.

In (14) it is not immediately clear that $R_{\underset{-}{i}\underset{+}{k}\underset{\cdot}{l}\underset{\cdot}{m};\underset{\cdot}{n}}$ is a tensor. We are now in a position to give a more useful form for (14) in which this is obvious.

$$\begin{aligned} R_{\underset{-}{i}\underset{+}{k}\underset{-}{l}\underset{+}{m};\underset{\cdot}{n}} + R_{\underset{-}{i}\underset{+}{k}\underset{+}{m}\underset{+}{n};\underset{\cdot}{l}} + R_{\underset{-}{i}\underset{+}{k}\underset{+}{n}\underset{-}{l};\underset{\cdot}{m}} &= R_{\underset{-}{i}\underset{+}{k}\underset{\cdot}{l}\underset{\cdot}{m};\underset{\cdot}{n}} - R_{ikrn}\Gamma_{ml}^{r} - R_{iklr}\Gamma_{mn}^{r} \\ &\quad - R_{ikrn}\Gamma_{ml}^{r} - R_{ikmr}\Gamma_{nl}^{r} - R_{iklr}\Gamma_{mn}^{r} - R_{iknr}\Gamma_{ml}^{r}. \end{aligned}$$

The first term on the right side of the equation vanishes by (14), the last six cancel out due to (15). Therefore,

$$R_{\underset{-}{i}\underset{+}{k}\underset{-}{l}\underset{+}{m};\underset{\cdot}{n}} + R_{\underset{-}{i}\underset{+}{k}\underset{+}{m}\underset{+}{n};\underset{\cdot}{l}} + R_{\underset{-}{i}\underset{+}{k}\underset{+}{n}\underset{-}{l};\underset{\cdot}{m}} = 0. \tag{14a}$$

4. The field-equations. We are now in a position to carry out the derivation of the identities for the field equations. In analogy to the gravitational theory, we contract (14a) by $g^{mi}g^{kl}$. (Note that the order of the indices is determined by the differentiation character of the corresponding indices in (14a).) Making use of (15), we get

$$g^{mi}g^{kl}[R_{\underset{-}{i}\underset{+}{k}\underset{-}{l}\underset{+}{m};\underset{\cdot}{n}} - R_{\underset{-}{i}\underset{+}{k}\underset{+}{n}\underset{+}{m};\underset{\cdot}{l}} - R_{\underset{-}{i}\underset{+}{k}\underset{-}{l}\underset{+}{n};\underset{\cdot}{m}}] = 0$$

or using (1a),

$$g^{kl}[g^{mi}R_{\underset{+}{i}\underset{-}{k}\underset{\cdot}{l}\underset{\cdot}{m}}]_{;\underset{\cdot}{n}} - g^{kl}[g^{mi}R_{\underset{+}{i}\underset{+}{k}\underset{\cdot}{n}\underset{\cdot}{m}}]_{;\underset{\cdot}{l}} - g^{mi}[g^{kl}R_{\underset{-}{i}\underset{-}{k}\underset{\cdot}{l}\underset{\cdot}{n}}]_{;\underset{\cdot}{m}} = 0. \tag{17}$$

Let us define

(18) $$R_{kl} = g^{mi}R_{iklm}$$

(19) $$S_{mi} = g^{kl}R_{iklm}$$

where

(18a) $$R_{kl} = g^{mi}g_{ai}R^{a}{}_{klm} = \delta_{a}{}^{m}R_{klm}{}^{a} = \Gamma_{kl}{}^{a}{}_{,a} - \Gamma_{ka}{}^{a}{}_{,l} - \Gamma_{ll}{}^{a}\Gamma_{ka}{}^{t} + \Gamma_{ta}{}^{a}\Gamma_{kl}{}^{t}.$$

Then we have

(17a) $$g^{kl}[R_{\underset{+}{k}\ \underset{-}{l}\ ;n} - R_{\underset{+}{k}\ \underset{+}{n}\ ;l} - S_{\underset{-}{n}\ \underset{-}{l}\ ;k}] = 0.$$

We need some connection between R and S. From (15), (16) we see that

$$R_{kiml} = \bar{R}_{iklm}.$$

Multiply by $g^{im}(= \bar{g}^{mi})$ and sum (i.e. contract):

(20) $$S_{lk} = \bar{R}_{kl}.$$

If R were Hermitian, R and S would be identical. Hence we have a new reason for requiring that R_{kl} should be Hermitian. But from (18a) we see that R_{kl} has an anti-Hermitian part (compare with (9)):

(21) $$\tfrac{1}{2}(R_{kl} - \bar{R}_{lk}) = \tfrac{1}{2}[(\Gamma_{al}{}^{a}{}_{,k} - \Gamma_{ka}{}^{a}{}_{,l}) - \Gamma_{kl}{}^{t}(\Gamma_{at}{}^{a} - \Gamma_{ta}{}^{a})].$$

From (10) we see that

(22) $$\Gamma_{la}{}^{a} = \frac{g_{,l}}{g} = (\tfrac{1}{2}\log|\det g_{ik}|)_{,l}.$$

Therefore,

$$\Gamma_{al}{}^{a}{}_{,k} - \Gamma_{ka}{}^{a}{}_{,l} = \Gamma_{\underset{\vee}{al}}{}^{a}{}_{,k} - \Gamma_{\underset{\vee}{ka}}{}^{a}{}_{,l}$$

(21a) $$\tfrac{1}{2}(R_{kl} - \bar{R}_{lk}) = -\tfrac{1}{2}(\Gamma_{\underset{\vee}{la}}{}^{a}{}_{,k} + \Gamma_{\underset{\vee}{ka}}{}^{a}{}_{,l} - \Gamma_{kl}{}^{t}\Gamma_{\underset{\vee}{ta}}{}^{a}).$$

From this we see that R_{kl} is Hermitian if we subject the field to the four conditions

(2) $$\Gamma_{\underset{\vee}{la}}{}^{a} = 0.$$

It then follows from (20) that

(20a) $$S_{lk} = R_{lk},$$

and (17a) becomes

(17b) $$g^{kl}[R_{\underset{+}{k}\ \underset{-}{l}\ ;n} - R_{\underset{+}{k}\ \underset{+}{n}\ ;l} - R_{\underset{-}{n}\ \underset{-}{l}\ ;k}] = 0.$$

These identities hold for all fields where Γ is defined by (1) and is subject to (2). We might jump to the conclusion that the field equations should stipulate the vanishing of all R_{kl}. This set, together with (1) and (2) would, however, be overdetermined. We can get a weaker set of equations by observing how $R_{\underset{\vee}{kl}}$ enters (17b). The contribution of $R_{\underset{\vee}{kl}}$ to the equations is:

$$g^{kl}[R_{\underset{+}{k}\ \underset{-}{l}\ ;n} - R_{\underset{+}{k}\ \underset{+}{n}\ ;l} - R_{\underset{-}{n}\ \underset{-}{l}\ ;k}]$$

which can be written as

$$
\begin{aligned}
g^{kl}[&R_{kl,n} - R_{sl}\Gamma_{kn}^{s} - R_{ks}\Gamma_{nl}^{s} - R_{kn,l} \\
&+ R_{sn}\Gamma_{kl}^{s} + R_{ks}\Gamma_{nl}^{s} - R_{nl,k} + R_{sl}\Gamma_{kn}^{s} + R_{ns}\Gamma_{kl}^{s}] \\
&= g^{kl}[R_{kl,n} - R_{kn,l} - R_{nl,k}] \\
&= g^{kl}[R_{kl,n} + R_{nk,l} + R_{ln,k}] \\
&= g^{kl}\,R_{kl,n}.
\end{aligned}
$$

Since we see that R_{kl} enters the equations only in the combination $R_{kl,n}$, it is natural to choose the field equations for R_{kl} as

(23) $$R_{kl,n} = 0$$

instead of $R_{kl} = 0$. So we get the field equations

(2) $$\Gamma_{is}^{s} = 0$$

(24) $$R_{kl} = 0$$

(23) $$R_{kl,n} = 0,$$

where the Γ_{ik}^{l} are defined by:

(1a) $$g_{i\,k;l} = 0.$$

The foregoing derivation shows how naturally we can extend general relativity theory to a non-symmetric field, and that the field-equations previously published are really the natural generalizations of the gravitational equations. If we were sure that a non-symmetric tensor g_{ik} is the right means for describing the structure of the generalized field, then we could hardly doubt that the above equations are the correct ones.

5. The variational principle. For comparison we include a derivation of the equations based on a variational principle. This is formally simpler than the previous derivation, but it has the disadvantage of making use of two apparently arbitrary restrictions of the $g - \Gamma$ field:

(2) $$\Gamma_{is}^{s} = 0,$$

(3) $$\mathfrak{g}^{is}_{,s} = 0.$$

On the other hand the equations (1) are deduced from the variation; we need not postulate them. It is advantageous to make use of Palatini's method in this derivation. As in sec. 3, we form the curvature tensor:

(11) $$R^{i}_{klm} = \Gamma_{kl}^{i}{}_{,m} - \Gamma_{km}^{i}{}_{,l} - \Gamma_{sl}^{i}\Gamma_{km}^{s} + \Gamma_{sm}^{i}\Gamma_{kl}^{s}.$$

By generalizing Palatini's method to the non-symmetric case, it is easy to verify that

(25) $$\delta R^{i}_{klm} = (\delta\Gamma_{k\,l}^{\underset{+-}{i}})_{;m} - (\delta\Gamma_{k\,m}^{\underset{++}{i}})_{;l}.$$

We choose the Hamiltonian function

(26) $$\mathfrak{H} = \mathfrak{g}^{kl}R_{kl}$$

(26a) $$\mathfrak{H} = \delta_{i}^{m}\mathfrak{g}^{kl}R^{i}_{klm}.$$

We vary (26a) relative to the Γ's:

$$\delta\mathfrak{H} = \delta_i^m \mathfrak{g}^{kl}(\delta R^i{}_{klm}) \tag{27}$$
$$= \delta_i^m \mathfrak{g}^{kl}[(\delta\Gamma_{k\,l\,\underset{+}{i}})_{;m} - (\delta\Gamma_{k\,m\,\underset{++}{i}})_{;l}].$$

For brevity we set

$$\mathfrak{A}^m = \delta_i^m \mathfrak{g}^{kl}(\delta\Gamma_{kl}{}^i) = \mathfrak{g}^{kl}(\delta\Gamma_{kl}{}^m) \tag{28}$$

$$\mathfrak{B}^l = \delta_i^m \mathfrak{g}^{kl}(\delta\Gamma_{km}{}^i) = \mathfrak{g}^{kl}(\delta\Gamma_{km}{}^m). \tag{29}$$

Then we can write (27) as

$$\delta\mathfrak{H} = \mathfrak{A}^m_{\underset{+}{;m}} - (\delta_{\underset{+}{i}}{}^m \mathfrak{g}^{k\,l}_{\underset{+}{+}})_{;m}(\delta\Gamma_{kl}{}^i)$$
$$- \mathfrak{B}^l_{\underset{-}{;l}} + (\delta_{\underset{+}{i}}{}^m \mathfrak{g}^{k\,l}_{\underset{+}{-}})_{;l}(\delta\Gamma_{km}{}^i).$$

We have to form the integral of $\delta\mathfrak{H}$. Let us see what $\mathfrak{A}^m_{\underset{+}{;m}}$ contributes to the integral. (See sec. 2.)

$$\mathfrak{A}^m_{\underset{+}{;m}} = \mathfrak{A}^m{}_{,m} + \mathfrak{A}^a \Gamma_{am}{}^m - \mathfrak{A}_{\underset{\underset{+}{m}\,a}{}} \tag{30}$$
$$= \mathfrak{A}^m{}_{,m} + \mathfrak{A}^m \Gamma_{\underset{\underset{+}{m}\,a}{}}.$$

The first term is an ordinary divergence, and hence contributes nothing to the integral. We see that we need (2) to make the second term vanish. By subjecting the field to (2) we make sure that $\mathfrak{A}^m_{\underset{+}{;m}}$ (and similarly $\mathfrak{B}^l_{\underset{-}{;l}}$) contributes nothing to the integral. So we may omit these from (27a) and write:

$$\delta\mathfrak{H} = [-(\delta_{\underset{+}{i}}{}^m \mathfrak{g}^{k\,l}_{\underset{+}{-}})_{;m} + (\delta_{\underset{+}{i}}{}^l \mathfrak{g}^{k\,m}_{\underset{+}{-}})_{;m}](\delta\Gamma_{kl}{}^i). \tag{27b}$$

Or since $\delta_{\underset{+}{i}}{}^l{}_{;m}$ vanishes:

$$\delta\mathfrak{H} = [-\mathfrak{g}^{k\,l}_{\underset{+}{-}\,;m} + \mathfrak{g}^{k\,m}_{\underset{+}{-}\,;m}\delta_{\underset{+}{i}}{}^l](\delta\Gamma_{kl}{}^i). \tag{27c}$$

We cannot conclude yet that the quantity in brackets vanishes, because the $\Gamma_{kl}{}^i$ are not independent but satisfy (2). But we could conclude the vanishing of these quantities if they depended on only 60 parameters instead of the 64 $\mathfrak{g}^{k\,l}_{\underset{+}{-}\,;\,\underset{+}{i}}$. This is actually so, for the following reason: we have

$$\tfrac{1}{2}(\mathfrak{g}^{k\,l}_{\underset{+}{-}\,;l} - \mathfrak{g}^{l\,k}_{\underset{+}{-}\,;l}) = \mathfrak{g}^{kl}_{\underset{\underset{+}{}}{},l} - \mathfrak{g}^{kl}_{\underset{\underset{+}{}}{}}\Gamma_{l\,a}{}^a. \tag{31}$$

By subjecting the field to (2) and (3), we make sure that these four quantities vanish. Hence only 60 of the $\mathfrak{g}^{k\,l}_{\underset{+}{-}\,;i}$ are independent. The same must be true of the square bracketed quantities in (27c). Thus we can conclude from (27c) that all these vanish. Contracting with respect to l and i we have $\mathfrak{g}^{k\,m}_{\underset{+}{-}\,;m} = 0$. Hence all the $\mathfrak{g}^{k\,l}_{\underset{+}{-}\,;i}$ vanish. Therefore also the $g_{\underset{+}{i}\,\underset{-}{k}\,;i}$. (See (1b), (1c).) Thus we have derived that

$$g_{\underset{+}{i}\,\underset{-}{k}\,;l} = 0. \tag{1a}$$

(It follows from these and (31) that either of the conditions (2) and (3) implies the other one.) We still have to vary (21) relative to $\mathfrak{g}^{ik}$. But we must remember that the $\mathfrak{g}^{ik}$ satisfy (3). This can be done most easily by setting

(32) $$\mathfrak{g}^{\underset{\sim}{ik}} = \mathfrak{g}^{ik\sigma}{}_{,\sigma}$$
$$\mathfrak{g}^{ik} = \mathfrak{g}^{\underset{\sim}{ik}} + \mathfrak{g}^{ik\sigma}{}_{,\sigma}$$

and varying with respect to $\mathfrak{g}^{\underset{\sim}{ik}}$ and $\mathfrak{g}^{ik\sigma}$, which are independent. ($\mathfrak{g}^{ik\sigma}$ is a tensor density antisymmetric in each pair of indices.) We get the equations

(23) $$R_{kl;\underset{\sim}{m}} = 0,$$

(24) $$R_{\underset{\sim}{kl}} = 0.$$

This completes the derivation of the field-equations.

We can further justify the *a priori* assumption of (2) by the fact that this equation is necessary and sufficient to make R_{kl} a Hermitian tensor. (See (21a).)

The Institute for Advanced Study
Princeton, N.J.